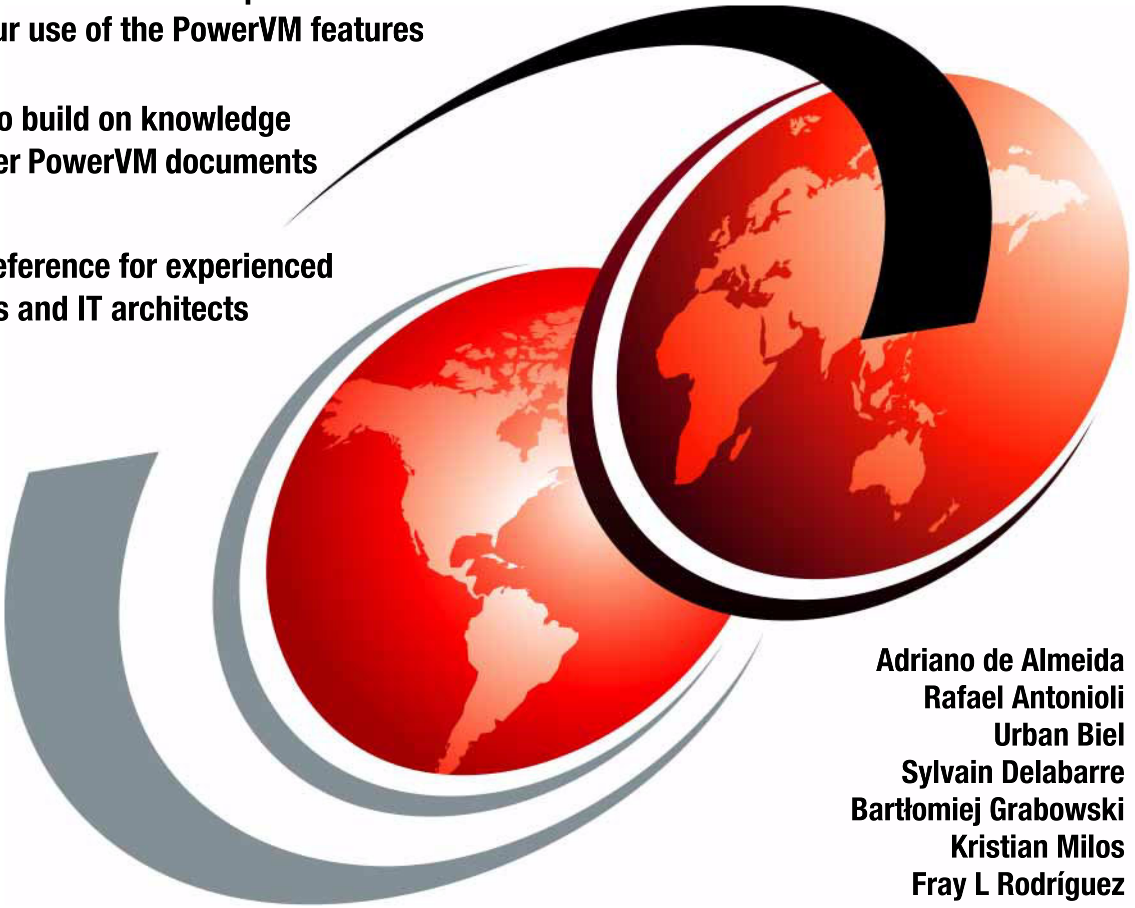
IBM PowerVM Best Practices

A collection of recommended practices to enhance your use of the PowerVM features

A resource to build on knowledge found in other PowerVM documents

A valuable reference for experienced IT specialists and IT architects



Adriano de Almeida
Rafael Antonioli
Urban Biel
Sylvain Delabarre
Bartłomiej Grabowski
Kristian Milos
Fray L Rodríguez

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM PowerVM Best Practices

October 2012

SG24-8062-00

Note: Before using this information and the product it supports, read the information in "Notices" on page xiii.

First Edition (October 2012)

This edition applies to:
PowerVM Enterprise Edition
Virtual I/O Server Version 2.2.1.4 (product number 5765-G34)
AIX Version 7.1 (product number 5765-G99)
IBM i Version 7.1 (product number 5770-SS1)
HMC Version 7.7.4.0 SP02
POWER7 System Firmware Version AL730_87

© Copyright International Business Machines Corporation 2012. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

© Copyright IBM Corp. 2012. All rights reserved.

Figures

© Copyright IBM Corp. 2012. All rights reserved.

Tables

© Copyright IBM Corp. 2012. All rights reserved.

Examples

© Copyright IBM Corp. 2012. All rights reserved.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:
This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

© Copyright IBM Corp. 2012. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Active Memory™
AIX®
BladeCenter®
DS4000®
DS8000®
Enterprise Storage Server®
Focal Point™
GDPS®
Geographically Dispersed Parallel Sysplex™
GPFS™
HACMP™
IBM®
iSeries®
Micro-Partitioning®
Parallel Sysplex®
PartnerWorld®
Power Systems™
POWER7®
PowerHA®
PowerVM®
POWER®
PureFlex™
PureSystems™
Redbooks®
Redpaper™
Redbooks (logo) ®
System p®
SystemMirror®
Tivoli®

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.

Preface

This IBM® Redbooks® publication provides best practices for planning, installing, maintaining, and monitoring the IBM PowerVM® Enterprise Edition virtualization features on IBM POWER7® processor technology-based servers.

PowerVM is a combination of hardware, PowerVM Hypervisor, and software, which includes other virtualization features, such as the Virtual I/O Server.

This publication is intended for experienced IT specialists and IT architects who want to learn about PowerVM best practices, and focuses on the following topics:

- Planning and general best practices
- Installation, migration, and configuration
- Administration and maintenance
- Storage and networking
- Performance monitoring
- Security
- PowerVM advanced features

This publication is written by a group of seven PowerVM experts from different countries around the world. These experts came together to bring their broad IT skills, depth of knowledge, and experiences from thousands of installations and configurations in different IBM client sites.

The team who wrote this book

This book was produced by a team of specialists from around the world, while working at the IBM International Technical Support Organization, Poughkeepsie Center.

Adriano de Almeida is an IT Specialist with IBM Integrated Technology Services Delivery in Brazil. He has worked at IBM for 13 years. His areas of expertise include IBM AIX®, PowerVM, IBM PowerHA®, and IBM HACMP™. He is a Certified Advanced Technical Expert on IBM System p®. He has worked extensively on PowerHA, PowerVM and AIX projects, health checking on IBM Power Systems™ environments and performing technical project leadership. He holds a degree in Computing Technology from the Faculdade de Tecnologia em Processamento de Dados do Litoral (FTPDL). He is also a coauthor of the Exploiting IBM PowerHA SystemMirror® Enterprise Edition IBM Redbooks publication.

© Copyright IBM Corp. 2012. All rights reserved.

Rafael Antonioli is a System Analyst at Banco do Brasil in Brazil. He has 12 years of experience with Linux and five years of experience in the AIX and PowerVM field. He holds a Master of Computer Science degree in Parallel and Distributed Computer Systems from Pontifical Catholic University of Rio Grande do Sul (PUCRS). His areas of expertise include implementation, support, and performance analysis of IBM PowerVM, IBM AIX, and IBM PowerHA.

Urban Biel is an IT Specialist in IBM Slovakia. He has been with IBM for six years. He holds a Master degree in Information Systems and Networking from Technical University of Kosice, Slovakia. His areas of expertise include Linux, AIX, PowerVM, PowerHA, IBM GPFS™ and also IBM enterprise disk storage systems. He has participated in several Redbooks publications.

Sylvain Delabarre is a certified IT Specialist at the Product and Solutions Support Center in Montpellier, France. He works as an IBM Power Systems Benchmark Manager. He has been with IBM France since 1988. He has 20 years of AIX System Administration and Power Systems experience working in service delivery, AIX, Virtual I/O Server, and HMC support for EMEA.

Bartłomiej Grabowski is an IBM iSeries® Senior Technical Specialist in DHL IT Services in the Czech Republic. He has seven years of experience with IBM i. He holds a Bachelor degree in Computer Science from Academy of Computer Science and Management in Bielsko-Biala. His areas of expertise include IBM i administration, PowerHA solutions that are based on hardware and software replication, Power Systems hardware, and IBM i virtualization that is based on PowerVM. He is an IBM Certified System Administrator. He was a coauthor of the IBM Active Memory™ Sharing Redbooks publication.

Kristian Milos is an IT Specialist at IBM Australia. Before working at IBM, he spent seven years working at the largest telecommunications organization in Australia. He has 10 years of experience working in enterprise environments, with the past six directly involved with implementing and maintaining AIX, PowerVM, PowerHA, and Power Systems environments.

Fray L Rodríguez is an IBM Consulting IT Specialist working in Power Systems Competitive Sales in the United States. He has 12 years of experience in IT and 17 years of experience in customer service. He holds a Bachelor degree in Software Engineering from the University of Texas at Dallas. Fray also holds 19 professional IBM Certifications, including IBM Expert Certified IT Specialist, IBM Technical Sales Expert on Power Systems, and IBM Advanced Technical Expert on Power Systems.

The project team that created this publication was managed by:

Scott Vetter, PMP
IBM Austin

Thanks to the following people for their contributions to this project:

Aaron Bolding, Thomas Bosworth, Ben Castillo, Shaival J Chokshi, Gareth Coates, Pedro Alves Coelho, Julie Craft, Rosa Davidson, Ingo Dimmer, Michael Felt, Rafael Camarda Silva Folco, Djamel Ghaoui, Chris Gibson, Chris Angel Gonzalez, Paranthaman Gopikaramanan, Randy Greenberg, Hemantha Gunasinghe, Margarita Hammond, Jimi Inge, Narutsugu Itoh, Chandrakant Jadhav, Robert C. Jennings, Anil Kalavakolanu, Bob Kovac, Kiet H Lam, Dominic Lancaster, Luciano Martins, Augie Mena, Walter Montes, Guérin Nicolas, Anderson Ferreira Nobre, Rajendra Patel, Viraf Patel, Thomas Prokop, Xiaohan Qin, Vani Ramagiri, Paisarn Ritthikidjaroenchai, Glenn Robinson, Björn Rodén, Humberto Roque, Morgan J Rosas, Stephane Saleur, Susan Schreitmueller, Jorge M Silvestre, Luiz Eduardo Simeone, Renato Stoffalette, Naoya Takizawa, Humberto Tadashi Tsubamoto, Morten Vaagmo, Tomaž Vincek, Richard Wale, Tom Watts, Evelyn Yeung, and Ken Yu.

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and client satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Introduction and planning

IBM PowerVM Best Practices, SG24-8062, provides a summary of best practices that are implemented by IBM PowerVM experts from around the world. PowerVM is available for use on all IBM Power Systems, including IBM POWER® technology-based IBM BladeCenter® servers and IBM PureSystems™.

PowerVM provides industrial-strength virtualization for IBM AIX, IBM i, and Linux operating systems. If there are any differences in best practices that pertain to these different operating systems, we point that out, as appropriate.

We suggest that you become familiar with the IBM Redbooks publications: *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, and *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590. We also suggest that you have some practical experience with virtualization before you use the material that is contained in this publication.

The IBM PowerVM Virtualization features are designed to meet all your enterprise production workload virtualization requirements. The development and test process for our Virtual I/O Server and virtual client partitions is as rigorous as it is for dedicated partitions. We also run our benchmarks in virtual environments. We suggest PowerVM for production, development, testing, or any other environment in which you would commonly use stand-alone servers or dedicated partitions.

© Copyright IBM Corp. 2012. All rights reserved.

This paper provides recommendations for the architecture, implementation, and configuration of PowerVM virtual environments. In areas where there are multiple configuration options, this publication provides recommendations that are based on the environment and the experiences of the authors and contributors.

It is possible that new best practices emerge as new hardware technologies become available.

Although this publication can be read from start to finish, it is written to enable the selection of individual topics of interest, and to go directly to them. The content is organized into seven major headings:

- Chapter 1, “Introduction and planning” on page 1, provides an introduction to PowerVM and best practices to plan and size your virtual environment.
- Chapter 2, “Installation, migration, and configuration” on page 15, describes best practices for the creation and installation of a Virtual I/O Server, and migration to new versions.
- Chapter 3, “Administration and maintenance” on page 31, covers best practices for daily maintenance tasks, backup, recovery, and troubleshooting of the Virtual I/O Server.
- Chapter 4, “Networking” on page 49, describes best practices for network architecture and configuration within the virtual environment.
- Chapter 5, “Storage” on page 61, covers best practices for storage architecture and configuration.
- Chapter 6, “Performance monitoring” on page 93, describes best practices for monitoring the Virtual I/O Server performance.
- Chapter 7, “Security and advanced IBM PowerVM features” on page 105, covers best practices for some of the more advanced PowerVM features.

1.1 Keeping track of PowerVM features

IBM is committed to the future of PowerVM, and is constantly enhancing current features and adding new ones to meet the needs of its clients.

For this reason, our first best practice is to become familiar with these PowerVM enhancements because they can enrich your environment. Table 1-1 provides a summary of these features at the time of writing, but you can continue tracking new features through the following website, along with other useful information about PowerVM:

http://ibm.com/systems/power/software/virtualization/features.html

Table 1-1 PowerVM features

PowerVM features	Description
PowerVM Hypervisor	Supports multiple operating environments on a single system.
Micro-Partitioning	Enables up to 10 virtual machines (VMs) per processor core.
Dynamic Logical Partitioning	Processor, memory, and I/O resources can be dynamically moved between VMs.
Shared Processor Pools	Processor resources for a group of VMs can be capped, reducing software license costs. VMs can use shared (capped or uncapped) processor resources. Processor resources can automatically move between VMs based on workload demands.
Shared Storage Pools	Storage resources for Power Systems servers and a Virtual I/O Server can be centralized in pools to optimize resource utilization.
Integrated Virtualization Manager	Simplifies VM creation and management for entry Power Systems servers and blades.
Live Partition Mobility	Live AIX, IBM i, and Linux VMs can be moved between frames, eliminating planned downtime.
Active Memory Sharing	Intelligently flows memory from one VM to another for increased memory utilization.
Active Memory Deduplication	Reduces memory consumption for Active Memory Sharing (AMS) configurations by detecting and eliminating duplicate memory pages.
NPIV	Simplifies the management and improves performance of Fibre Channel SAN environments.

Note: Table 1-1 on page 3 shows all the PowerVM features, but these features are grouped in three edition packages: Express, Standard, and Enterprise, to best meet your virtualization needs. The following website shows the features that ship with each PowerVM edition:

http://www.ibm.com/systems/power/software/virtualization/editions.html

1.2 Virtual I/O Server specifications

The following section defines the range of configuration possibilities, which includes the minimum number of resources that are needed and the maximum number of resources allowed.

1.2.1 Virtual I/O Server minimum requirements

Table 1-2 provides the minimum resources that are needed to create a Virtual I/O Server. Consider these values when you are planning and designing your Virtual I/O Server.

Table 1-2 Minimum resources that are required for a Virtual I/O Server

Resource	**Requirement**
Hardware Management Console (HMC) or Integrated Virtualization Manager (IVM)	The HMC or IVM is required to create the logical partition and assign resources.
Storage adapter	The server logical partition needs at least one storage adapter.
Physical disk	The disk must be at least 30 GB.
Ethernet adapter	If you want to route network traffic from virtual Ethernet adapters to a Shared Ethernet Adapter (SEA), you need an Ethernet adapter.
Memory	For POWER7 processor-based systems, at least 768 MB of memory is required.
Processor	At least 0.1 processor is required.
PowerVM Editions	One of the three PowerVM Editions: Express, Standard, or Enterprise.

1.2.2 Configuration considerations of the Virtual I/O Server

Consider the following general configuration design points:

- Virtual Small Computer System Interface (SCSI) supports the following connection standards for backing devices: Fibre Channel, SCSI, SCSI Redundant Array of Independent Disks (RAID), IP-based SCSI (iSCSI), serial-attached SCSI (SAS), Serial Advanced Technology Attachment (SATA), Universal Serial Bus (USB), and Integrated Device Electronics (IDE).
- The SCSI protocol defines mandatory and optional commands. While the virtual SCSI supports all of the mandatory commands, not all of the optional commands are supported.
- There might be utilization implications when you use virtual SCSI devices. Because the client/server model is made up of layers of function, the use of virtual SCSI can use more processor cycles when it is processing I/O requests.
- The Virtual I/O Server is a dedicated logical partition, to be used only for Virtual I/O Server operations. Third-party applications are supported by the vendors that produced them.
- If there is a resource shortage, performance degradation might occur. If a Virtual I/O Server is serving many resources to other logical partitions, ensure that enough processor power is available. If there is a high workload across virtual Ethernet adapters and virtual disks, logical partitions might experience delays in accessing resources.
- Logical volumes and files that are exported as virtual SCSI disks are always configured as single path devices on the client logical partition.
- Logical volumes or files that are exported as virtual SCSI disks that are part of the *root volume group* (rootvg) are not persistent if you reinstall the Virtual I/O Server. However, they are persistent if you update the Virtual I/O Server to a new service pack. Therefore, before you reinstall the Virtual I/O Server, ensure that you back up the corresponding virtual disks of those virtual clients. When you export logical volumes, it is best to export logical volumes from a volume group other than the root volume group. When you export files, it is best to create the file storage pools and a virtual media repository in a parent storage pool other than the root volume group.
- Only Ethernet adapters can be shared. Other types of network adapters cannot be shared (at the time of writing).
- IP forwarding is not supported on the Virtual I/O Server.
- The maximum number of virtual adapters can be any value in the range 2 - 65536. However, if you set the maximum number of virtual adapters to a value higher than 1024, the logical partition might fail to activate.

Furthermore, server firmware might require more system memory to manage the virtual adapters.

- The Virtual I/O Server supports client logical partitions that run IBM AIX 5.3 or later, IBM i 6.1 or later, and Linux. The minimum version of Linux that is supported might be different depending on the hardware model. Check the hardware manual for the model you need to confirm the minimum version you can use.

1.2.3 Logical Volume Manager limits in a Virtual I/O Server

If you plan to use logical volumes as backing devices, consider the following limitations for the Logical Volume Manager (LVM) components that are used in a Virtual I/O Server. Table 1-3 shows these limitations.

Table 1-3 Limitations for storage management

Category	Limit
Volume groups	4096 per system
Physical volumes	1024 per volume group
Physical partitions	1024 per volume group
Logical volumes	1024 per volume group
Logical partitions	No limit

1.3 Planning your Virtual I/O Server environment

PowerVM offers great flexibility when you configure a Virtual I/O Server, and with that flexibility, there are a number of choices you can make. This flexibility means that there is more than one correct way to design and implement your virtual environment.

Based on this flexibility, it is important for you to illustrate a picture of the environment by listing all the possible correct options for implementing each scenario. Also include the consequences of each choice. You can make an informed decision by using this practice. Reaching best practices for your environment are a matter of choosing the option that best fits the workloads you plan to run. What is considered a best practice in a lab, might not be a best practice for an enterprise banking server.

Section 1.3 describes the choices that you encounter to help you plan your environment.

1.3.1 System Planning Tool

The System Planning Tool (SPT) is a browser-based application that helps you plan, design, and validate system configurations. It is a best practice to use this tool to create your initial setup and to distribute the hardware resources among the dedicated and virtual partitions. This tool has many functions and many built-in checks to assist you with your system planning.

It is also a good practice to use this tool to plan your partitions. This practice is beneficial because you can deploy the system plans that are generated with the SPT. You can use these plans to create the partition profiles on your Hardware Management Console (HMC) or the Integrated Virtualization Manager (IVM).

Another benefit of using the SPT is that it can help you calculate how much of the system memory that the Power Hypervisor uses depending on your configuration. This way, you can plan according to how much remains available for use in your partitions. Finally, it is also a best practice to print the SPT validation report to ensure that you are meeting all the system requirements. The SPT is at this website:

http://ibm.com/systems/support/tools/systemplanningtool

1.3.2 Document as you go

Do not wait until the end of the project to document your environment. Whether you prefer to use a spreadsheet, a database, or any other type of software, write down all the requirements you have for your environment, your objectives, your choices, and your solutions. Then update it as you plan, design, and build your virtual environment.

1.3.3 Hardware planning

If you are involved in the decision making process of purchasing the IBM Power Systems server that is used to build your virtual environment, this is a good time to start filling out your documentation spreadsheet. List all the functional and nonfunctional requirements that you have in your environment to ensure that the system is configured with all the hardware components, software, and PowerVM features that you need.

Think of accessibility, availability, scalability, performance, security, compliance, disaster recovery, and so on, and everything you need to do to meet the requirements of these categories. For example, you might decide that you want to use N-Port ID Virtualization (NPIV). Then you need to ensure that you purchase 8-Gb adapters, which are a hardware requirement. Another requirement might be that you are able to move your virtual clients from one physical server to another without any downtime. This feature is called *Live Partition Mobility*, and is included in the IBM PowerVM Enterprise Edition. Therefore, you need to include this requirement to ensure that Enterprise Edition is included in the frame, as opposed to the Standard or Express Edition.

Power Systems servers offer outstanding flexibility as to the number of adapters, cores, memory, and other components and features that you can configure in each server. Therefore, discuss your requirements with your IBM technical sales team, and they can configure a server that meets your needs.

Other hardware features

When you are planning your hardware configuration, remember that there are other features that are available in Power Systems servers that are not part of PowerVM. However, these other features can be useful in your virtual environment.

For example, Active Memory Expansion (AME) is a feature that can be activated with any Power Systems server and can be used to increase memory capacity up to 100%. This feature can be used with the PowerVM feature AMS, or independently.

There are other features, such as Active Memory Mirroring (AMM), that are only available on Enterprise class servers. AMM can provide mirroring of the Power Hypervisor code among different dual inline memory modules (DIMMs). This feature enhances the availability of a server in case a DIMM failure occurs in one of the DIMMs that hold the Hypervisor code.

As a best practice, get a list of all the optional Reliability, Availability, and Serviceability (RAS) features that are available in the Power Systems server that you purchase. Plan to implement the features.

1.3.4 Sizing your Virtual I/O Server

Sizing a Virtual I/O Server can be a challenging task for many IT Architects and IT Specialists. In fact, it can be argued that the only way to truly size any server is to run it with a real workload, monitor it, and tune it accordingly.

The following guidelines are intended to help you get started with the sizing of a Virtual I/O Server environment. The Virtual I/O Servers offer great flexibility, such as the option to run different virtual clients with different operating systems and thousands of independent software vendor (ISV) applications. Therefore, it is important that you monitor the performance of the Virtual I/O Server and adjust its resources, if needed.

As a best practice, have at least two Ethernet adapters and two Fibre Channel adapters in your Virtual I/O Server.

For small Power Systems servers, with eight cores or less, you can start with 0.5 of entitled processor capacity and 2 GB of memory.

For more powerful Power Systems servers, we use the example of a Virtual I/O Server that uses two 1 Gb Ethernet adapters and two 4 Gb Fibre Channel adapters, supporting around 20 virtual clients. For this scenario, use one core for the entitled processor capacity and 4 GB of entitled memory when you run AIX or Linux. If the virtual clients are running IBM i and are using virtual SCSI, use 1.5 for entitled processor capacity and 4 GB of entitled memory. If the IBM i virtual client is using NPIV, one core of entitled capacity is a good starting point.

High speed adapters, such as 10 Gb Ethernet and 8-Gb Fibre Channel, require more memory for buffering. For those environments, use 6 GB of memory in each scenario, regardless of the operating system that is running on the virtual I/O client. This is also especially important to consider if you plan to use NPIV. It is important because the Virtual I/O Server conducts work that is similar to a virtual storage area network (SAN) switch that is passing packets back and forth between your SAN and the virtual I/O clients. Therefore, the Virtual I/O Server requires more memory for each virtual Fibre Channel adapter that is created in a virtual client.

Estimating an exact amount of memory that is needed per adapter is difficult without using tools to look at the specific workloads. It is difficult because the requirements of the adapters vary depending on their technologies, Peripheral Component Interconnect-X (PCI-X) and PCI Express (PCIe), and their configurations. A rule of thumb for estimating the amount of memory that is needed by each adapter, is to add 512 MB for each physical high speed Ethernet adapter in the Virtual I/O Server. This number also includes the memory that is needed for virtual Ethernet adapters in the virtual clients. Also, add 140 MB for each virtual Fibre Channel adapter in the client.

Based on this example, for a Virtual I/O Server that has two 10 Gb Ethernet adapters and two 8 Gb Fibre Channel adapters, supports 20 virtual clients, and assumes that each client has one Virtual Fibre Channel adapter, we have the following scenario:

2 GB for base workload + 1 GB (512 MG x 2) for Ethernet adapters + 2.8 GB (20 x 140 MB) for Virtual Fibre Channel adapters = 5.8 GB. Therefore, our general recommendation is to have 6 GB of memory.

Core speeds vary: Remember that core speeds vary from system to system, and IBM is constantly improving the speed and efficiency of the IBM POWER processor cores and memory. Therefore, the referenced guidelines are, again, just a starting point. When you create your Virtual I/O Server environment, including all the virtual clients, be sure to test and monitor it to ensure that the assigned resources are appropriate to handle the load on the Virtual I/O Server.

1.3.5 IBM Systems Workload Estimator

If you want a more precise method than what is described in the guidelines in section 1.3.4, “Sizing your Virtual I/O Server” on page 8, use the IBM Systems Workload Estimator (WLE) tool. WLE is a web-based sizing tool, which allows you to size Power Systems servers, IBM System X, and IBM PureFlex™ Systems. For information about this tool and to obtain a download version, see this website:

`http://ibm.com/systems/support/tools/estimator`

For the Virtual I/O Server, the WLE sizing is based on metrics that drive activity through it. When sizing, consider metrics such as the number and size of the disk I/Os, and the number and size of the network I/Os.

1.3.6 Shared or dedicated resources

During the planning session, decide whether your Virtual I/O Server and virtual client partitions are to use dedicated resources, processors and memory, or shared resources.

If you use shared processors and shared memory, you get better utilization, and you can often add new workloads or increase existing ones without more resources. This sharing translates into lower core-based software licensing costs, therefore, saving money.

Some people prefer to have a dedicated amount of processing units and memory for a production partition. If you choose to use dedicated resources in each partition, your sizing of the environment needs to consider peak times during the day. Peak times include backup time or peak seasons, such as back to school and holidays, which might require you to have unused resources during the rest of the year.

For planning purposes, whichever option you choose, it is a good practice to understand how these options affect your environment. Also consider the amount of resources that you need to satisfy your workloads.

1.3.7 Single, dual, or multiple Virtual I/O Servers

IBM Power Systems servers are resilient. Therefore, a single Virtual I/O Server is sufficient to support many small environments, but redundancy is always a plus. Therefore, having dual or multiple Virtual I/O Servers is a best practice for production environments.

The key here is to understand what your availability requirements are to decide what you need. Do you need redundancy at the hardware level or are you implementing redundancy at the application level across multiple logical partitions or physical frames? Do you have any Service Level Agreements (SLAs) that require your servers to be up 99.999% of the time? Are you able to schedule maintenance once or twice a year to do updates? Are there any penalties for planned or unplanned downtime?

These are the type of questions you need to answer to plan whether you get a single or dual Virtual I/O Server, or even multiple Virtual I/O Servers. For instance, some environments have heavy amounts of network and storage I/O traffic. Therefore, some administrators prefer to have four Virtual I/O Servers: Two to handle the network traffic and two to handle the I/O from the SAN. Besides availability, the idea behind this type of configuration is to improve performance.

Mission critical workloads have different requirements to development or to test partitions. Therefore, as a best practice, plan for the right number of resources that are needed by one or multiple Virtual I/O Servers, depending on the workloads you plan to run.

1.3.8 Network and storage components

Now you can also plan on the number and type of network and storage components you need in your Virtual I/O Server. For instance, if you plan to have dual Virtual I/O Servers, you might ask yourself how many Ethernet adapters you need and at what speed? Will Etherchannel be used? Will you use shared Ethernet adapter failover from one Virtual I/O Server to the other? Are multiple Virtual LANs (VLANs) needed to separate the network traffic for security purposes?

You need to think the same way about your storage component needs. Will NPIV or virtual SCSI be used? If you use virtual SCSI, will Shared Storage Pools be used? Will the boot occur from internal disk drives or SAN?

Think about all these components in advance to determine the best way to configure and implement them. Remember to document your choices and the reasons behind them. Table 1-4 contains some of the network and storage components you might consider.

Table 1-4 Network and storage components

Network components	**Storage components**
Physical Ethernet adapters	Physical Fibre Channel adapters
Virtual Ethernet adapters	Virtual Fibre Channel adapters
Shared Ethernet adapters	Virtual SCSI adapters
Backup virtual Ethernet adapters	SAN volumes
Virtual LANs (VLANs)	Multi-path I/O
Etherchannel or Link Aggregation devices	Mirroring internal drives
	Storage pools (volume groups)
	Shared Storage Pools

1.3.9 Slot numbering and naming conventions

The use of a slot numbering system and naming convention, and sticking to it, can save a great deal of time when troubleshooting issues. This system is also helpful when new employees are learning how everything is mapped from the Virtual I/O Server to the virtual clients. Therefore, plan to create a numbering system that suits your needs, and follow it.

With any of the slot numbering options, it is best if you map the same slot number on the Virtual I/O Server with the same slot number in the virtual client. Assuming you use that approach, you can use the following ideas to help create a numbering method.

One easy option for a dual Virtual I/O Server environment is to use even numbers for one server and odd numbers for the other. In this case, one Virtual I/O Server has 11, 13, 15, and so on, while the second Virtual I/O Server has 12, 14, and so on.

If you also want to identify the virtual client through the slot number without having to look at the mapping, you can add its partition ID to the odd or even number. The following example shows this process:

```
<virtual_client_ID><(virtual_server_odd or Virtual_server_even)>
```

For example, in a system with dual Virtual I/O Servers with IDs 1 and 2, and a virtual client with ID 3, we use odd numbers for Virtual I/O Server 1, as in 31, 33, 35, and so on. The second Virtual I/O Server uses even numbers, as in 32, 34, 36, and so on.

This system might not be appropriate though if you have many clients, because the slot number is too high. Also, consider that when you use odd and even numbers, it is difficult to maintain a system if you are frequently using Live Partition Mobility (LPM) to move virtual clients from one physical system to another.

Use low slot numbers: It is a best practice to use low slot numbers and define only the number of virtual adapters that you use. This practice is preferred because the Power Hypervisor uses a small amount of memory for each virtual device that you define.

A different method for slot numbering is to use adapter ID = XY, where the following values apply:

X= adapter type, using 1 for Ethernet, 2 for Fibre Channel, and 3 for SCSI

Y= next available slot number

With this system, your virtual Ethernet adapters are 11 - 19. Your Virtual Fibre Channel is 20 - 29, and your virtual SCSI is 30 - 39.

LPM: When you use LPM, if a target system does not have the slot available that you want to use, it asks if it can use a new slot number. If slot numbers change, it is important that you keep current documentation of your environment.

For naming convention methods for the virtual adapters in a Virtual I/O Server, use a name that is descriptive of the virtual client that uses this virtual adapter.

1.3.10 Operating systems on virtual client partitions

A Virtual I/O Server can host client partitions that run AIX, IBM i, or Linux. Therefore, it is important to consider the operating system that is running in the virtual client when you plan your Virtual I/O Server configuration. For instance, some tuning options might be more beneficial on one virtual client operating system than another. Throughout the book, we mention such differences as they arise.

You can also use the Fix Level Recommendation Tool (FLRT) to ensure that the version of the Virtual I/O Server is compatible with other components. The version needs to be compatible with the firmware on the server, the version of the operating system on the virtual client, the HMC or IVM, and other IBM software that might be in your environment. To use the FLRT, see this website:

http://www14.software.ibm.com/webapp/set2/flrt/home

2

Installation, migration, and configuration

This chapter describes the best practices of installing, migrating, and configuring your Virtual I/O Server. We also address the creation of the Virtual I/O Server profile.

© Copyright IBM Corp. 2012. All rights reserved.

2.1 Creating a Virtual I/O Server profile

Section 2.1 describes the best practices to create a Virtual I/O Server profile. To create a profile, you have the options of using the Hardware Management Console (HMC), the Integrated Virtualization Management (IVM), and the System Planning Tool (SPT). For ease of deployment of multiple virtual partitions, we suggest you use the SPT to create and deploy your partitions.

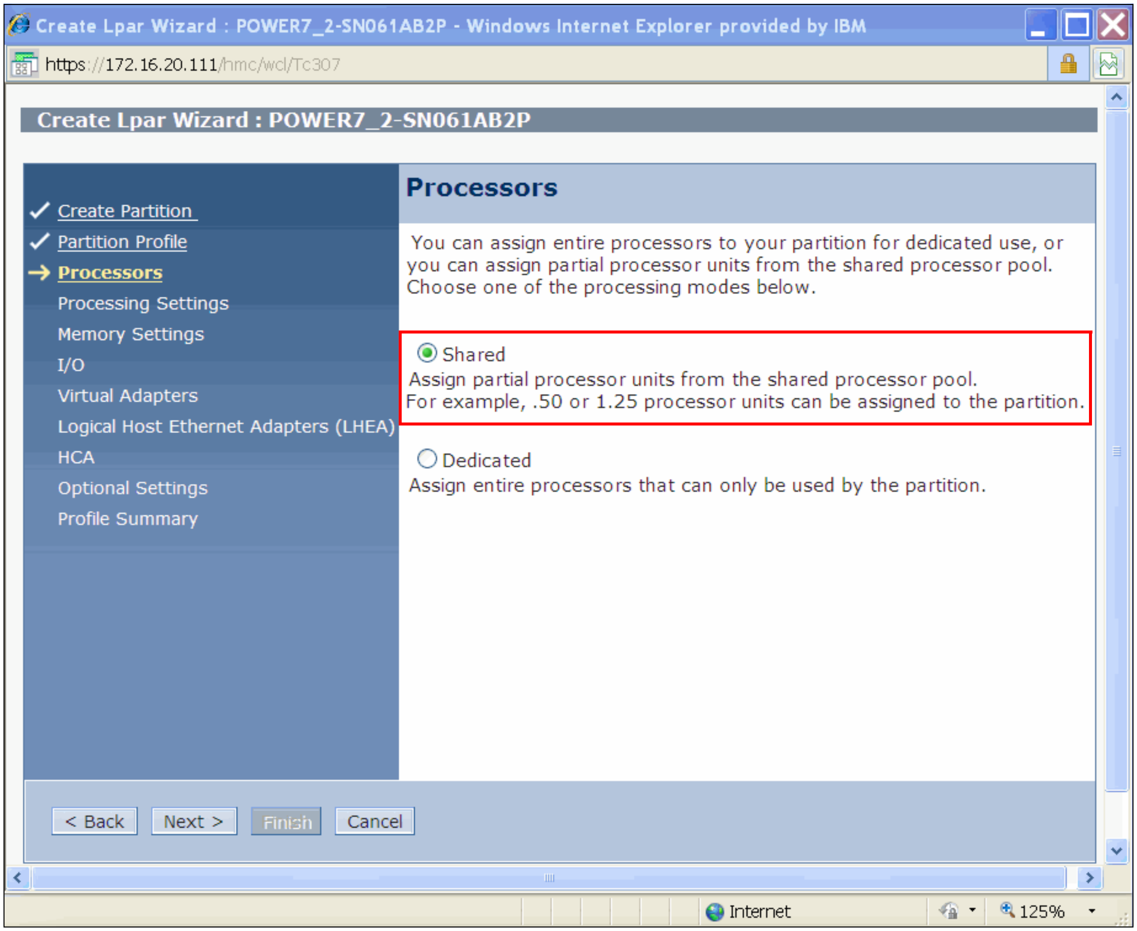
The following subsections are best practices to create the Virtual I/O Server and virtual client profiles, regardless of the management tool that is used to create them.

The complete set of instructions for creating a Virtual I/O Server is described in the Redbooks publication, *PowerVM Virtualization Introduction and Configuration* SG24-7940.

2.1.1 Processing mode - shared or dedicated

The Power Hypervisor manages the allocation of processor resources to partitions within a system. Each partition is allocated processor resources according to its profile. The processor allocation can be dedicated to that partition or shared with other partitions within a processor pool.

It is a best practice to use the shared processing mode in a Virtual I/O Server because of the changes in the virtual client workloads. Figure 2-1 on page 17 shows the selection of the shared processing mode.


Create Lpar Wizard : POWER7_2-SN061AB2P - Windows Internet Explorer provided by IBM
https://172.16.20.111/hmc/wcl/Tc307
Create Lpar Wizard : POWER7_2-SN061AB2P
✓ Create Partition
✓ Partition Profile
→ Processors
Processing Settings
Memory Settings
I/O
Virtual Adapters
Logical Host Ethernet Adapters (LHEA)
HCA
Optional Settings
Profile Summary
< Back
Next >
Finish
Cancel
Internet
125%

Processors

You can assign entire processors to your partition for dedicated use, or you can assign partial processor units from the shared processor pool. Choose one of the processing modes below.

● Shared
Assign partial processor units from the shared processor pool.
For example, .50 or 1.25 processor units can be assigned to the partition.

○ Dedicated
Assign entire processors that can only be used by the partition.

Figure 2-1 Select processor type

2.1.2 Processing settings

When you are defining the values for physical and virtual processor units, you need to assign a minimum, a wanted, and a maximum value.

Physical processing units

For minimum processing units, the best practice is to set this value to 0.1. This number is the lowest configured unit value that can be assigned to a partition, and which allows the Virtual I/O Server to boot.

For wanted processing units, follow the sizing considerations that are described in 1.3.4, “Sizing your Virtual I/O Server” on page 8. Set this value to a number that meets the needs of the estimated workload.

For maximum processing units, round up the wanted processing units value to the next whole number, and add 50%. For example, if the wanted value is 1.2, the maximum value is 3.

It is important to allow room between the wanted and the maximum processing units. This suggestion is because you can only increase the wanted value dynamically, through a dynamic LPAR operation, up to the maximum value. At the same time, it is important not to set the maximum value too high because the Power Hypervisor uses more memory the higher it is.

Virtual processing units

It is important to remember that there is Hypervisor CPU usage that is associated with the number of online virtual processors. Therefore, carefully consider the capacity requirements of the Virtual I/O Server before you make a decision.

Uncapped partitions require virtual processors so they can exceed their entitlement. However, do not make these values too high because of the Hypervisor CPU usage and because if there is not enough processing capacity, it might cause excessive context switching. This switching might, in turn, cause performance degradation. Below are some general best practices to follow:

- For minimum virtual processors, round up the minimum processing units value to the next whole number. For example, if the minimum processing units are 0.1, the minimum virtual processors value is 1.
- For wanted virtual processors, add 50% to the wanted processing units value, and round it up to a whole number. For example, if the wanted number of processing units is set to 1, the wanted number of virtual processors is 2.
- For maximum virtual processors, add 50% to the wanted virtual processors. For example, if the number of wanted virtual processors is 2, the maximum value is 3.

Capped or uncapped

The sharing mode of processor cores can be set to capped or uncapped. Capped partitions have a preset amount of maximum processing unit entitlement. However, partitions that are configured with uncapped processor resources are able to use all of their allocation, plus any unused processing units in a shared processor pool.

The load on the Virtual I/O Server varies depending on the demands of the virtual clients. Therefore, you might see spikes in processor or memory usage throughout the day. To address these changes in workload, and to achieve better utilization, use shared and uncapped processors. Therefore, uncapping can provide a significant benefit to partitions that have spikes in utilization.

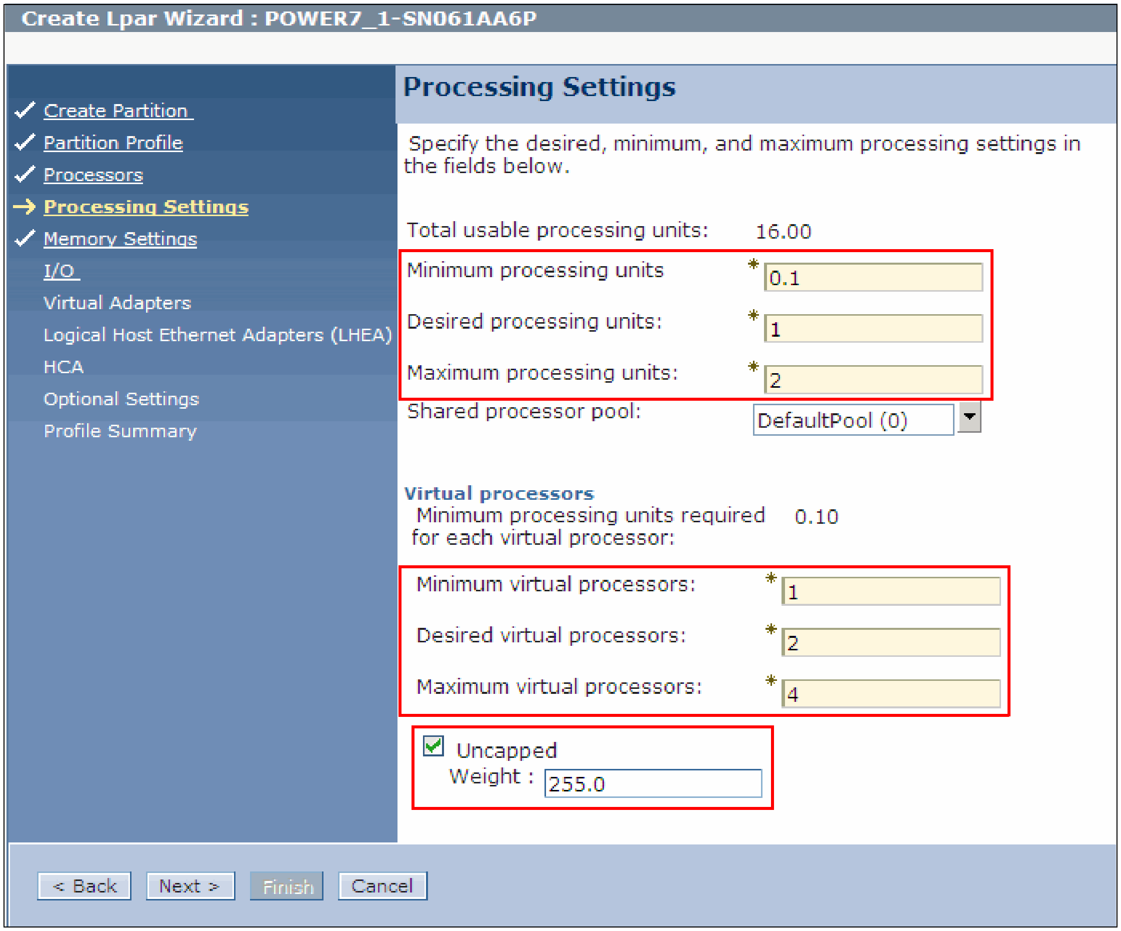
Weight

When you choose the uncapped partition option, you also need to choose a weight value. For the Virtual I/O Server, best practice is to configure a weight value higher than the virtual clients. The maximum configured value is 255. The Virtual I/O Server must have priority to the processor resources in the frame. Table 2-1 shows an example of distributing weight among different types of environments, depending on their importance.

Table 2-1 Processor weighting example

Value	**Usage**
255	Virtual I/O Server
200	Production
100	Pre-production
50	Development
25	Test

Figure 2-2 on page 20 shows an example of a best practice configuration of Processing Settings.



Processing Settings

Specify the desired, minimum, and maximum processing settings in the fields below.

Total usable processing units: 16.00

Minimum processing units * 0.1

Desired processing units: * 1

Maximum processing units: * 2

Shared processor pool: DefaultPool (0)

Virtual processors

Minimum processing units required for each virtual processor: 0.10

Minimum virtual processors: * 1

Desired virtual processors: * 2

Maximum virtual processors: * 4

☑ Uncapped

Weight : 255.0

Figure 2-2 Processing Settings best practices

2.1.3 Memory settings

When you define the values for physical memory, you need to assign a minimum, a wanted, and a maximum value.

For wanted memory, follow the sizing suggestions that are described in 1.3.4, “Sizing your Virtual I/O Server” on page 8. Use these sizes to meet the workload demands of your Virtual I/O Server. From this value, we can determine the minimum and maximum memory values.

For minimum memory, use 50% of the wanted memory value. For example, if the wanted value is 4 GB, the minimum value is 2 GB.

For maximum memory, add 50% to the wanted memory value. For example, if the wanted value is 4 GB, the maximum value is 6 GB.

Similarly to processing units, it is important to allow room between the wanted and the maximum memory values. This consideration is because you can only increase the wanted value dynamically, through a dynamic LPAR operation, up to the maximum value.

The maximum memory value is also the number that is used when you calculate the amount of memory that is needed for the page tables to support this partition. For this reason, it is not advantageous to set this maximum setting to an unreasonably high amount. This recommendation is because it would waste memory by setting memory aside for page tables that the partition does not need.

Amount of memory: The SPT can help you estimate the amount of memory that is required by the Power Hypervisor depending on your IBM Power Systems server and its configuration. Also, the HMC can show you the amount of memory that is available for partition use under the frame properties, and the amount that is reserved for the Hypervisor.

2.1.4 Physical I/O adapters

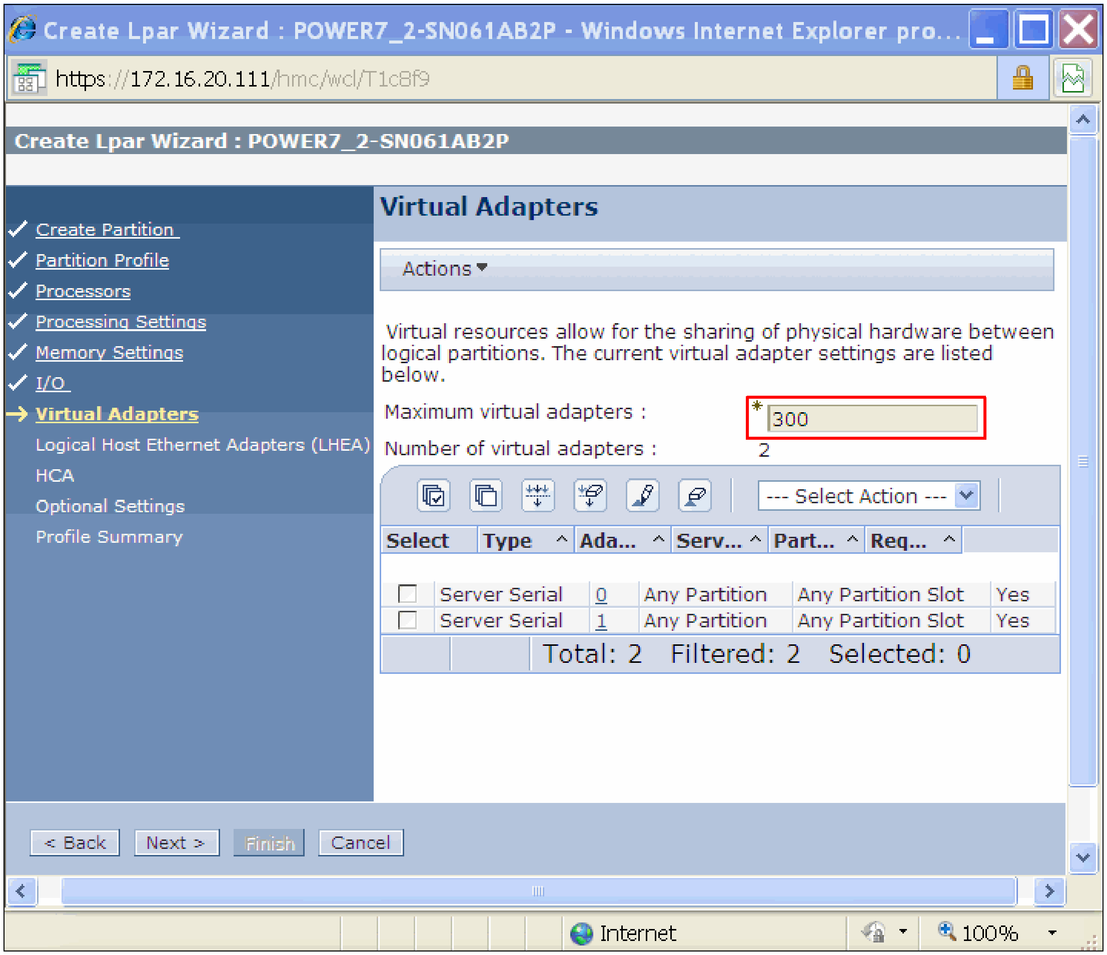
It is a best practice to configure all the physical I/O adapters to the wanted memory value. Having them set to the required value, prevents dynamic logical partition (DLPAR) operations from working. The aim is to allow adapter changes to happen dynamically with no downtime required on the partition. Also, an adapter which is set to *required* that is in a failed or removed state, prevents the LPAR from activating.

2.1.5 Virtual I/O adapters

It is a good practice to have a good planning naming process in the Virtual I/O Server because managing a large environment can be complex.

Maximum virtual adapters

Before you set the maximum virtual adapter limit, you need to consider the number of slots that your virtualized environment uses. The larger you set the maximum virtual adapter value, the more memory the Hypervisor needs to reserve to manage them. Figure 2-3 on page 22 shows the process for setting the maximum number of virtual adapters.



Virtual Adapters

Virtual resources allow for the sharing of physical hardware between logical partitions. The current virtual adapter settings are listed below.

Maximum virtual adapters : * 300

Number of virtual adapters : 2

Select	Type	Ada...	Serv...	Part...	Req...
☐	Server Serial	0	Any Partition	Any Partition Slot	Yes
☐	Server Serial	1	Any Partition	Any Partition Slot	Yes

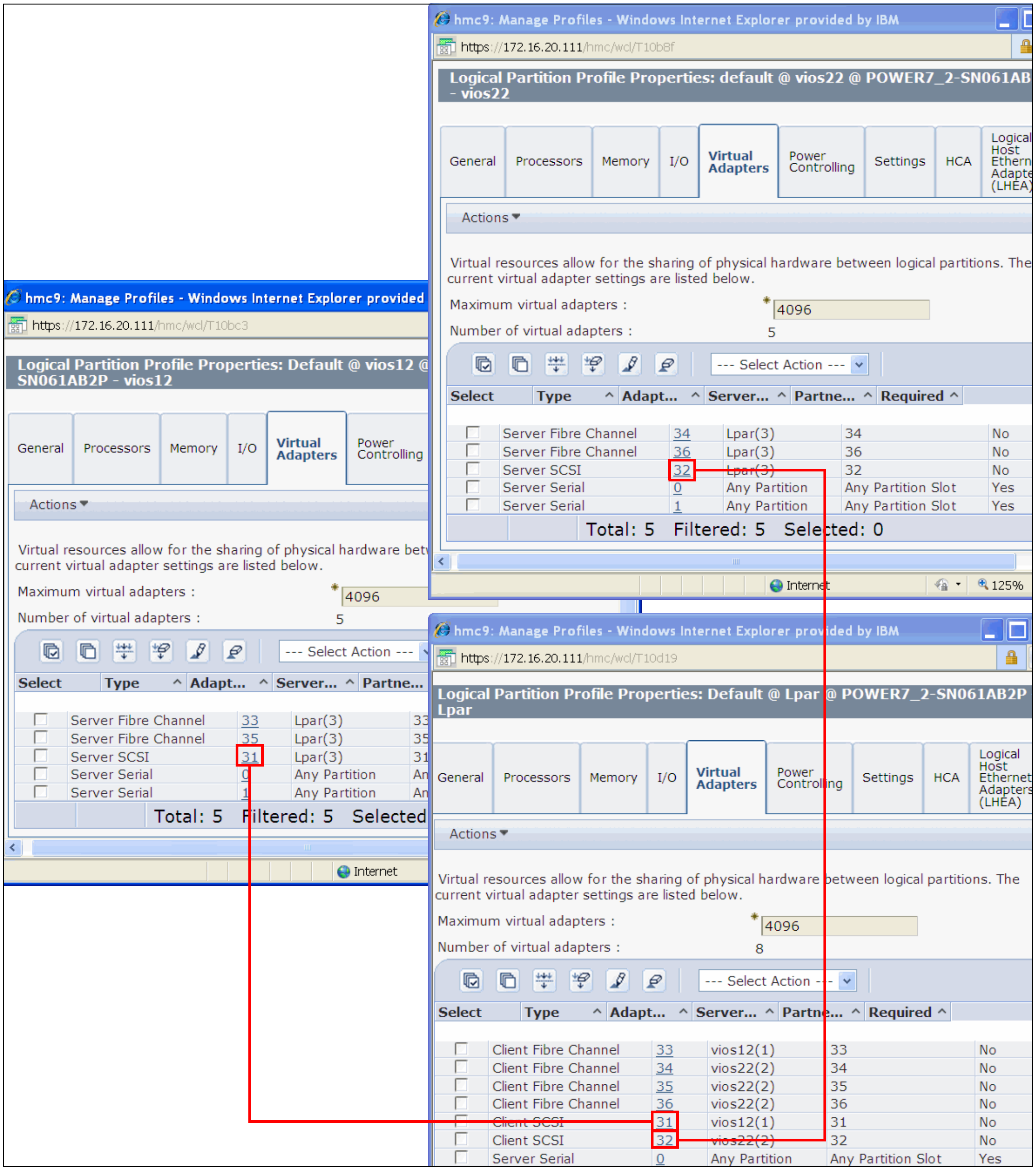
Total: 2 Filtered: 2 Selected: 0

Figure 2-3 Setting the maximum number of virtual adapters

Virtual adapter slot numbers

When you configure the virtual SCSI and virtual Fibre Channel adapters, it is a best practice to keep the virtual adapter IDs on the client and the server, the same. This practice can be useful when you manage and trace your storage configuration.

In Figure 2-4 on page 23, we show an example of the virtual adapter relationship between two Virtual I/O Servers and one virtual I/O client.



Logical Partition Profile Properties: Default @ vios12 @ SN061AB2P - vios12

Virtual resources allow for the sharing of physical hardware bet current virtual adapter settings are listed below.

Maximum virtual adapters : * 4096

Number of virtual adapters : 5

Select	Type	Adapt...	Server...	Partne...
□	Server Fibre Channel	33	Lpar(3)	33
□	Server Fibre Channel	35	Lpar(3)	35
□	Server SCSI	31	Lpar(3)	31
□	Server Serial	0	Any Partition	An
□	Server Serial	1	Any Partition	An

Total: 5 Filtered: 5 Selected

hmc9: Manage Profiles - Windows Internet Explorer provided by IBM
https://172.16.20.111/hmc/wcl/T10b8f
General
Processors
Memory
I/O
Virtual Adapters
Power Controlling
Settings
HCA
Logical Host Ethern Adapte (LHEA)
Actions
--- Select Action ---
Internet
125%

Logical Partition Profile Properties: default @ vios22 @ POWER7_2-SN061AB - vios22

Virtual resources allow for the sharing of physical hardware between logical partitions. The current virtual adapter settings are listed below.

Maximum virtual adapters : * 4096

Number of virtual adapters : 5

Select	Type	Adapt...	Server...	Partne...	Required
□	Server Fibre Channel	34	Lpar(3)	34	No
□	Server Fibre Channel	36	Lpar(3)	36	No
□	Server SCSI	32	Lpar(3)	32	No
□	Server Serial	0	Any Partition	Any Partition Slot	Yes
□	Server Serial	1	Any Partition	Any Partition Slot	Yes

Total: 5 Filtered: 5 Selected: 0

hmc9: Manage Profiles - Windows Internet Explorer provided by IBM
https://172.16.20.111/hmc/wcl/T10d19
General
Processors
Memory
I/O
Virtual Adapters
Power Controlling
Settings
HCA
Logical Host Ethernet Adapters (LHEA)
Actions
--- Select Action ---

Logical Partition Profile Properties: Default @ Lpar @ POWER7_2-SN061AB2P Lpar

Virtual resources allow for the sharing of physical hardware between logical partitions. The current virtual adapter settings are listed below.

Maximum virtual adapters : * 4096

Number of virtual adapters : 8

Select	Type	Adapt...	Server...	Partne...	Required
□	Client Fibre Channel	33	vios12(1)	33	No
□	Client Fibre Channel	34	vios22(2)	34	No
□	Client Fibre Channel	35	vios22(2)	35	No
□	Client Fibre Channel	36	vios22(2)	36	No
□	Client SCSI	31	vios12(1)	31	No
□	Client SCSI	32	vios22(2)	32	No
□	Server Serial	0	Any Partition	Any Partition Slot	Yes

Figure 2-4 Partition profile properties for source and target virtual adapters

Live Partition Mobility: If you are planning to implement Live Partition Mobility, set all virtual adapters to *wanted*. Required virtual I/O adapters prevent Live Partition Mobility operations.

2.1.6 Deploying a Virtual I/O Server with the System Planning Tool

After you configure your logical partitions in the System Planning Tool, you can save your system plan file (.sysplan) and import it into an HMC or IVM. You can then deploy the system plan to one or more frames. System plan deployment delivers the following benefits to you:

- You can use a system plan to partition a frame and deploy partitions without having to re-create the partition profiles. This plan saves time and reduces the possibility of errors.
- You can easily review the partition configurations within the system plan, as necessary.
- You can deploy multiple identical systems, almost as easily as a single system.
- You can archive the system plan as a permanent electronic record of the systems that you create.

2.2 Virtual I/O Server installation

Described here are the best practices to install the Virtual I/O Server. In general, we suggest that you use the Network Installation Manager (NIM) in your environment for installations, updates, and fixes. This suggestion applies to the Virtual I/O Server, too. If you are going to install more than one Virtual I/O Server, the NIM helps you significantly. Following is a basic installation procedure:

- Install the Virtual I/O Server by using the NIM. For details on how to create NIM resources for the Virtual I/O Server installation, see this website:

  http://www.ibm.com/support/docview.wss?uid=isg3T1011386#4

- Do not perform any virtualization configurations, and do not mirror the root volume group (rootvg).
- Install all necessary device drivers, apply the updates and fixes, and then reboot.
- Perform a backup of the Virtual I/O Server to create a golden image.
- Use the golden image to deploy all the other Virtual I/O Servers.

If you do not have a NIM in your environment, install it from a DVD, and then apply the basic installation procedures. You can use the **alt_root_vg** command to deploy other Virtual I/O Servers.

alt_root_vg command: If you boot from a cloned disk that is made by the **alt_root_vg** command, we suggest you remove obsolete devices that are in the defined state. Also, you might need to reconfigure the Reliable Scalable Cluster Technology (RSCT) subsystem by using the **recfgct** command. For more information about the **recfgct** command, see this website:
http://www.ibm.com/support/docview.wss?uid=swg21295662

If you boot the Virtual I/O Server from the local Small Computer System Interface (SCSI) disks, remember to mirror the rootvg by using the **mirrorios** command. And, set the boot list with both the SCSI disks by using the **bootlist** command.

2.3 Updating fix packs, service packs, and interim fixes

In addition to fix packs, the Virtual I/O Server might also have service packs, depending on the number of changes that are needed. Virtual I/O Server Service Packs consist of critical changes that are found between fix pack releases.

2.3.1 Virtual I/O Server service strategy

To ensure the reliability, availability, and serviceability of the Virtual I/O Server, we suggest that you update it to the latest available release, fix pack, or service pack. The latest level contains all of the cumulative fixes for the Virtual I/O Server. Interim fixes are only provided at the latest level of the Virtual I/O Server. You need to upgrade to the latest service level or fix pack to install an interim fix.

Updating strategy

The following suggestions assist you in determining the best update strategy for your enterprise:

- Set a date, every six months, to review your current firmware and software patch levels.
- Verify the suggested code levels by using the Fix Level Recommendation Tool (FLRT) on the IBM Support Site:

  http://www.software.ibm.com/webapp/set2/flrt/home

- Check for the Virtual I/O Server release lifecycles to plan your next upgrade. See this website:

  http://www.ibm.com/software/support/lifecycleapp/PLCSearch.wss?q=%28Virtual+I%2FO+Server%29+or++%28PowerVM%29&scope=&ibm-view-btn.x=3&ibm-view-btn.y=9&sort=S

- There are several options for downloading and installing a Virtual I/O Server update, such as downloading ISO images, packages, or installing from optical media. To check the latest release and instructions for Virtual I/O Server fix updates, see IBM Fix Central at this website:

  http://www.ibm.com/support/fixcentral/

> **ISO images:** Do not use utilities to extract ISO images of Virtual I/O fix packs or service packs to a local directory or NFS mount point. Burn these images to media. If you need fix packs or service packs on a local directory, download them as a package.

- Ensure that a regular maintenance window is available to conduct firmware updates and patching. Once a year is the suggested time frame to conduct the updates.
- When you do system firmware updates from one major release to another, always update the HMC to the latest available version first, along with any mandatory HMC patches. Then, do the firmware updates. If the operating system is being updated as well, update the operating system first, then the HMC code, and lastly the system firmware.
- In a dual HMC configuration always update both HMCs in a single maintenance window. Or, disconnect one HMC until it is updated to the same level as the other HMC.

2.3.2 Approved third-party applications in the Virtual I/O Server

It is a best practice to check third-party application lists that are allowed to be installed in the Virtual I/O Server. Installing applications that are not listed as a valid application on the IBM PartnerWorld® website, invalidate the IBM Virtual I/O Server support. See this website:

http://dbluedst.pok.ibm.com/DCF/isg/isgintra.nsf/all/T1010620

IBM support teams do not support these applications. For problems with these applications, contact your appropriate vendor support.

2.3.3 Applying fix packs, service packs, and interim fixes

Before you begin installing your applications, confirm that the following statements are true:

- The managed frame of the HMC is attached to the system and both the managed frame and HMC are working properly.
- When you update your Virtual I/O Server, do it in a console window (instead of a telnet) or run Secure Shell (SSH) to the Virtual I/O Server. For example, on AIX, run the script file name to keep a log, run SSH to connect to the HMC, then run `vtmenu`.
- Always check the readme instructions before you apply your updates.
- If you do not have a recent backup, run the **`backupios`** command and keep the output file in a safe location.

  You can use the **`alt_root_vg`** command to clone your rootvg to an alternate disk and update the Virtual I/O Server to the next fix pack level; see Example 2-1. This command can be done without taking the system down for an extended period and mitigating an outage risk.

  This cloning can be done by creating a copy of the current rootvg on an alternate disk and simultaneously applying the fix pack updates. If needed, the **`bootlist`** command can be run after the new disk is booted. The bootlist can be changed to boot back to the older level of the operating system in the event of an issue. Example 2-1 shows the **`alt_root_vg`** command.

Example 2-1 Using the **`alt_root_vg`** *command*

```
$ alt_root_vg -target hdisk1 -bundle update_all -location /mnt
```

Before you shut down the Virtual I/O Server, follow these steps:

- On a single Virtual I/O Server environment, shut down the virtual I/O clients that are connected to the Virtual I/O Server. Or disable any virtual resource that is in use.
- In a dual Virtual I/O Server environment, check that the alternate Virtual I/O Server is up and running and is serving I/O to the client. (Shared Ethernet Adapter (SEA) failover, virtual SCSI mapping, virtual Fibre Channel mapping, and so on).
  - If there is Logical Volume Manager mirroring on clients, check that both disks of any mirrored volume group are available in the system and the mirroring is properly synchronized.

  - If there is a SEA failover, check the configuration of the priority, and that the backup is active on the second Virtual I/O Server.
  - If there is a Network Interface Backup (NIB), check that the Etherchannel is configured properly on the virtual I/O clients.

2.4 Virtual I/O Server migration

The purpose of migration is to change the Virtual I/O Server software to a new version while you preserve your configuration files. For example, the migration process saves configuration files, installs the new software on your system, and then restores your configuration.

2.4.1 Options to migrate the Virtual I/O Server

You can migrate the Virtual I/O Server in the following ways:

- Migrate the Virtual I/O Server using the NIM.
- Migrate the Virtual I/O Server from the HMC.
- Migrate the Virtual I/O Server from optical media.

The best practice is to migrate the Virtual I/O Server using the NIM. Detailed information about each of the migration options is on the IBM Power Systems Hardware Information Center. See this website:

http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp

2.4.2 Virtual I/O Server migration considerations

Before you start, confirm that the following statements are true:

- The managed system of the HMC is attached to the system and both the managed system and the HMC are working properly.
- If you do not have a recent backup, run the **backupios** command and keep the output file in a safe location.

> **Cloning:** You can use the **alt_root_vg** command to clone your Virtual I/O Server. Cloning the running rootvg allows for the creation of a backup copy of the root volume group. This copy can be used as a backup in case the rootvg fails.

- Note the location of the rootvg disks.
- Confirm that you have the Migration DVD for the Virtual I/O Server instead of the Installation DVD for the Virtual I/O Server.

> **Installation media:** The media for the migration and for the installation are different. Using the installation media overwrites your current Virtual I/O Server configuration.

Before you shut down the Virtual I/O Server that you want to migrate, we suggest you check the following scenarios:

- In a single Virtual I/O Server configuration, during the migration, the client partition needs to be shut down. When the migration is complete, and the Virtual I/O Server is restarted, the client partition can be brought up without any further configuration.
- In a dual Virtual I/O Server environment, you can migrate one Virtual I/O Server at a time to avoid any interruption of service to the clients:
  - If there is Logical Volume Manager mirroring on clients, check that both disks of any mirrored volume group are available in the system and the mirroring is properly synchronized.
  - If there is Shared Ethernet Adapter (SEA) failover, check the configuration of the priority, and that the backup is active on the second Virtual I/O Server.
  - If there is a Network Interface Backup (NIB), check that the Etherchannel is configured properly on the virtual I/O clients.

> **Folding:** Processor folding currently is not supported for Virtual I/O Server partitions. If folding is enabled on your Virtual I/O Server and migration media is used to move from Virtual I/O Server 1.5 to 2.1.0.13 FP 23, or later, processor folding remains enabled. Upgrading by using migration media does not change the processor folding state. If you installed Virtual I/O Server 2.1.3.0, or later, and did not change the folding policy, then folding is disabled.

2.4.3 Multipathing software

When you plan to migrate your Virtual I/O Server, you must check that the multipathing driver software is supported on the new version.

For example, when you migrate the Virtual I/O Server from version 1.x to 2.x, you must migrate the Subsystem Device Driver (SDD) or the SSD Path Control Module (SDDPCM) software. Detailed information is on the following website:

http://www.ibm.com/support/docview.wss?uid=ssg1S7002686

For third-party multipathing software, consult the vendor documentation to ensure compatibility with the Virtual I/O Server version.

3

Administration and maintenance

This chapter describes best practices for general administration topics of the virtual I/O environment, with a focus on backing up and restoring the Virtual I/O Server. We also address dynamic logical partition operations, network installation manager resilience, and the virtual media repository.

Virtual I/O clients depend on the Virtual I/O Server for services. Therefore, it is critical that the entire virtual I/O environment is backed up, and that system restore and startup procedures include the Virtual I/O Server.

© Copyright IBM Corp. 2012. All rights reserved.

3.1 Backing up and restoring the Virtual I/O Server

The Virtual I/O Server, as is the case of any other critical server within an IT environment, needs to be backed up as part of a data recovery program for an enterprise. Section 3.1 sets out a strategy to coordinate the backup of the Virtual I/O Server into your current backup strategy.

This section also describes a complete solution that can be used to restore the Virtual I/O Server to another system, independent of the system type or model number. If you want to conduct a backup of just the Virtual I/O Server, only a subset of Section 3.1 is required.

3.1.1 When to back up the Virtual I/O Server

A complete disaster recovery strategy for the Virtual I/O Server includes backing up the following components that make up a virtualized environment. A change to any one of the following devices requires a new backup:

- External device configuration, for example, storage area network (SAN) and storage subsystem configuration.
- Memory, processor, virtual, and physical devices.
- The Virtual I/O Server operating system.
- User-defined virtual devices that couple the virtual and physical environments. This configuration can be considered *virtual device mappings*, or *metadata*.

You might notice that there is no mention of the operating system, installed applications, or application data of the virtual clients that are listed. There are no references because the Virtual I/O Server manages only the devices and the linking of these devices along with the Virtual I/O Server operating system itself. The virtual clients that are running IBM AIX, IBM i, or Linux need to have a backup strategy that is independently defined as part of your existing server backup strategy.

For example, if you have an AIX virtual client that is made up of virtual disks and a virtual network, you would still have a **`mksysb`** and **`savevg.`** Or you have an equivalent strategy in place to back up the system. This backup strategy can rely on the virtual infrastructure. For example, this strategy includes backing up to an IBM Tivoli® Storage Manager (TSM) server over a virtual network interface through a physical Shared Ethernet Adapter (SEA).

3.1.2 Virtual I/O Server backup strategy

The following section defines best practices to perform backup operations.

External device configuration

If a natural or man-made disaster destroys a complete site, planning for that occurrence can be included into the end-to-end backup strategy. This planning is part of your current disaster recovery (DR) strategy. The backup strategy depends on the hardware specifics of the storage, networking equipment, and SAN devices, to name a few. Examples of the type of information that is needed to record, include the virtual local area network (VLAN) or logical unit number (LUN) information from a storage subsystem.

This recording information is beyond the scope of this document. However, we mention it here to make you aware that a complete DR solution for a physical or virtual server environment has a dependency on this information. The method to collect and record the information depends not only on the vendor and model of the infrastructure systems at the primary site, but also what is present at the DR site.

Resources that are defined on the HMC and IVM

The definition of the Virtual I/O Server logical partition on the *Hardware Management Console* (HMC) and *Integrated Virtualization Manager* (IVM) includes, for example, how much processor and memory allocation is needed. Also consider what physical adapters are used. In addition to this consideration, you have the virtual device configuration (for example, virtual Ethernet adapters and what VLAN IDs to which they belong) that needs to be captured. The backup and restore of this data is beyond the scope of this document. For more information, see these IBM Redbooks publications:

- *Hardware Management Console V7 Handbook*, SG24-7491
  http://www.redbooks.ibm.com/abstracts/sg247491.html?Open
- *Integrated Virtualization Manager on IBM System p5*, REDP-4061
  http://www.redbooks.ibm.com/abstracts/redp4061.html?Open

Consider that you might need to rebuild selected HMC and IVM profiles from scratch on new hardware, especially if you are planning for disaster recovery. In this case, it is important to have detailed documentation of the configuration, such as how many Ethernet cards and Fibre adapters are needed. Using the System Planning Tool (SPT) to create system plans can help you record such information.

Backing up the Virtual I/O Server

The Virtual I/O Server operating system consists of the base code, fix packs, custom device drivers to support disk subsystems, and user-defined customization. An example of user-defined customization can be as simple as the changing of the Message of the Day or the security settings.

These settings, after an initial setup, will probably not change aside from the application of fix packs. Therefore, a sensible backup strategy for the Virtual I/O Server is needed before you apply the fix packs or make configuration changes. Section 3.1.3, “Backing up user-defined virtual devices” on page 35 covers this strategy. However, having a daily backup of the virtual and logical configuration can save time in restoring from configuration errors.

There are three ways to back up the Virtual I/O Server:

- Back up to tape.
- Back up to DVD-RAM.
- Back up to a remote file.

Backing up the Virtual I/O Server to a remote file is the most common and best practices method.

backupios command

The **`backupios`** command performs a backup of the Virtual I/O Server to a tape device, an optical device, or a file system (local file system or a remotely mounted Network File System (NFS)).

Backing up to a remote file system satisfies having the Virtual I/O Server backup at a remote location. It also allows the backup to be restored from either a Network Installation Management (NIM) server or the HMC. In Example 3-1, a Virtual I/O Server backup is done to an NFS mount which resides on a NIM server.

Example 3-1 Virtual I/O Server backup to a remote file system

```
$ mount nim:/export/vios_backup /mnt
$ backupios -file /mnt -nomedialib
Backup in progress.  This command can take a considerable amount of
time to complete, please be patient...
```

> **`-nomedialib` flag:** The **`-nomedialib`** flag excludes the contents of the virtual media repository from the backup. Unless explicitly required, excluding the repository significantly reduces the size of the backup.

The **backupios** command that is used in Example 3-1 on page 34, creates a full backup tar file package named *nim_resources.tar*. This package includes all of the resources that are needed to restore a Virtual I/O Server (mksysb image, `bosinst.data`, network boot image, and the Shared Product Object Tree (SPOT)) from a NIM or HMC by using the **installios** command. Section 3.1.4, “Restoring the Virtual I/O Server” on page 38, describes the restoration methods.

Best practice dictates that a full backup is taken before you make any configuration changes to the Virtual I/O Server. In addition to the full backup, a scheduled weekly backup is also a good practice. You can schedule this job by using the **crontab** command.

viosbr *command*

The **viosbr** command performs a backup of the Virtual I/O Server virtual and logical configuration. In Example 3-2, a scheduled backup of the virtual and logical configuration is set up, and existing backups are listed. The backup frequency is daily, and the number of backup files to keep is seven.

Example 3-2 Daily viosbr schedule

```
$ viosbr -backup -file vios22viosbr -frequency daily -numfiles 7
Backup of this node (vios22) successful
$ viosbr -view -list
vios22viosbr.01.tar.gz
```

At a minimum, backing up the virtual and logical configuration data before you make changes to the Virtual I/O Server, can help in recovering from configuration errors.

3.1.3 Backing up user-defined virtual devices

The **backupios** command backs up the Virtual I/O Server operating system, but more than that is needed to rebuild a server:

- If you are restoring to the same server, some information might be available such as data structures (storage pools, volume groups, and logical volumes) that are held on non-root volume group (rootvg) disks.
- If you are restoring to new hardware, these devices cannot be automatically recovered because the disk structures do not exist.
- If the physical devices exist in the same location and structures such that the logical volumes are intact, the virtual devices such as the virtual Small Computer System Interface (SCSI), virtual Fibre Channel, and SEAs are recovered during the restoration.

In a DR situation where these disk structures do not exist and network cards are at different location codes, you need to ensure that you back up the following devices:

- Any user-defined disk structures such as storage pools or volume groups and logical volumes.
- The linking of the virtual device through to the physical devices.

These devices are mostly created at the Virtual I/O Server build and deploy time, but change depending on when new clients are added or changes are made.

Backing up disk structures

Use the **savevgstruct** command to back up user-defined disk structures. This command writes a backup of the structure of a named volume group (and therefore, storage pool) to the /home/ios/vgbackups directory.

The **savevgstruct** command is automatically called before the backup commences for all active non-rootvg volume groups or storage pools on a Virtual I/O Server when the **backupios** command is run. Because this command is called before the backup commences, the volume group structures are included in the system backup. For this reason, you can use the **backupios** command to back up the disk structure as well.

Activate volume groups: The volume group needs to be activated for the backup to succeed. Only the active volume groups or storage pools are automatically backed up by the **backupios** command. Use the **lsvg** or **lssp** commands to list and **activatevg** to activate the volume groups or storage pools if necessary before you start the backup.

Backing up device mappings

The physical and virtual device mappings are contained in the **viosbr** backup that is described in “viosbr command” on page 35. You can use the **viosbr** command to view the device mappings, as shown in Example 3-3.

Example 3-3 List device mappings from a **viosbr** *backup*

```
$ viosbr -view -file vios22viosbr.01.tar.gz -mapping

Details in: vios22viosbr.01
SVSA                Physloc                            Client Partition ID
------------------- ---------------------------------- --------------------
vhost0              U8233.E8B.061AB2P-V2-C30           0x00000003

VTD                      rootvg_lpar01
Status                   Available
```

```
LUN                      0x8100000000000000
Backing Device           hdisk3
Physloc                  U78A0.001.DNWHZS4-P2-D6
Mirrored                 false

SVEA    Physloc
------- ---------------------------------------
ent6    U8233.E8B.061AB2P-V2-C111-T1

VTD                      ent11
Status                   Available
Backing Device           ent10
Physloc                  U78A0.001.DNWHZS4-P1-C6-T2
```

Slot numbers: It is also vitally important to use the slot numbers as a reference for the virtual SCSI, virtual Fibre Channel, and virtual Ethernet devices. Do not use the vhost/vfchost number or ent number as a reference.

The vhost/vfchost and ent devices are assigned by the Virtual I/O Server as they are found at boot time or when the **cfgdev** command is run. If you add in more devices after subsequent boots or with the **cfgdev** command, these devices are sequentially numbered.

The important information in Example 3-3 on page 36, is not vhost0, but that the virtual SCSI server in slot 30 (the C30 value in the location code) is mapped to physical volume hdisk3.

In addition to the information that is stored in the **viosbr** backup file, a full system configuration backup can be captured by using the **snap** command. This information enables the Virtual I/O Server to be rebuilt from the installation media if necessary. The crucial information in the **snap** is the output from the following commands:

- Network settings
  - **netstat -state**
  - **netstat -routinfo**
  - **netstat -routtable**
  - **lsdev -dev entX -attr**
  - **cfgnamesrv -ls**
  - **hostmap -ls**
  - **optimizenet -list**
  - **entstat -all entX**

- Physical and logical volume devices
  - **lspv**
  - **lsvg**
  - **lsvg -lv VolumeGroup**
- Physical and logical adapters
  - **lsdev -type adapter**
- Code levels, users, and security
  - **ioslevel**
  - **motd**
  - **loginmsg**
  - **lsuser**
  - **viosecure -firewall view**
  - **viosecure -view -nonint**

We suggest that you gather this information in the same time frame as the previous information.

The `/home/padmin` directory (which contains the **snap** output data) is backed up using the **backupios** command. Therefore, it is a good location to collect configuration information before a backup.

> **snap** output**:** Keep in mind that if a system memory dump exists on the Virtual I/O Server, it is also captured in the **snap** output.

3.1.4 Restoring the Virtual I/O Server

Similarly to backups, there are a few different methods in restoring the Virtual I/O Server. The most common and best practice is to restore from either the HMC or NIM.

Restoring the Virtual I/O Server from the HMC

If you made a full backup to the file, as shown in "backupios command" on page 34 (not a **mksysb** backup, but a full backup that creates the `nim_resources.tar` file), you can use the HMC to restore the Virtual I/O Server by using the **installios** command.

The tar file needs to be available on either the HMC, a DVD, or an NFS share. In this scenario, we use the preferred NFS method. Assuming that the directory that holds the `nim_resources.tar` file is exported from an NFS server, you now log on to the HMC with a suitable user ID. Then, run the **installios** command. Example 3-4 on page 39 shows this command.

`installios` command: The trailing slash in the NFS location `nim:/export/vios_backup/` must be included in the command as shown.

The configure client network interface setting must be disabled, as shown by the `-n` option. This step is necessary because the physical adapter in which we are installing the backup, might already be used by a SEA. If so, the IP configuration fails. Log in and configure the IP if necessary after the installation by using a console session.

A definition of each command option is available in the **`installios`** man page.

Example 3-4 The installios command from the HMC

```
hscroot@hmc9:~> installios -p vios22 -i 172.16.22.33 -S 255.255.252.0
-g 172.16.20.1 -d 172.16.20.41:/export/vios_backup/ -s
POWER7_2-SN061AB2P -m 00:21:5E:AA:81:21 -r default -n -P auto -D auto
...
...Output truncated
...
# Connecting to vios22
# Connected
# Checking for power off.
# Power off complete.
# Power on vios22 to Open Firmware.
# Power on complete.
# Client IP address is 172.16.22.33.
# Server IP address is 172.16.20.111.
# Gateway IP address is 172.16.20.1.
# Subnetmask IP address is 255.255.252.0.
# Getting adapter location codes.
# /lhea@200000000000000/ethernet@200000000000002 ping successful.
# Network booting install adapter.
# bootp sent over network.
# Network boot proceeding, lpar_netboot is exiting.
# Finished.
```

Now open a terminal console on the server to which you are restoring, in case user input is required.

Tip: If the **`installios`** command seems to be taking a long time to restore, this lag is most commonly caused by a speed or duplex misconfiguration in the network.

Restoring the Virtual I/O Server using a NIM server

The **installios** command is also available on the NIM server, but now, it only supports installations from the base media of the Virtual I/O Server. The method that we used from the NIM server is to restore the **mksysb** image. This image can be the **mksysb** image that is generated with the **-mksysb** flag in the **backupios** command. Or, you can extract the **mksysb** image from the `nim_resources.tar` file.

Whichever method you use to obtain the **mksysb**, you need to register the **mksysb** as a NIM resource and generate a SPOT from the **mksysb**. Do this registration after you have the backup image on the NIM server. With those prerequisites satisfied, you can now prepare the system for the restore. Example 3-5 shows this process.

Example 3-5 Preparing NIM resources for Virtual I/O Server restore

```
# nim -o define -t mksysb \
-a server=master \
-a location=/export/vios_backup/vios22.mksysb vios22_mksysb
# nim -o define -t spot \
-a server=master \
-a location=/export/vios_backup/spot \
-a source=vios22_mksysb vios22_spot
# nim -o bos_inst \
-a source=mksysb \
-a mksysb=vios22_mksysb \
-a spot=vios22_spot \
-a installp_flags=-agX \
-a no_nim_client=yes \
-a boot_client=no \
-a accept_licenses=yes vios22
```

> **Important:** The `no_nim_client=yes` option instructs the NIM server to not register the Virtual I/O Server as a NIM client. When this option is set to *no*, the NIM server configures an IP address on the physical adapter which was used during installation. If this adapter is apart from the SEA, it causes errors during boot.

With all the NIM resources ready for installation, you can now proceed to starting the Virtual I/O Server logical partition in SMS mode and perform a network boot. Further information about this process can be found in *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590.

Restoring disk structure and device mappings

If you restore a Virtual I/O Server to the same logical partition, all devices and links to those devices are restored. However, if you are restoring to another logical partition or system, these device mappings fail. This failure occurs because the system serial number makes up part of the virtual device location code; therefore, it is not present during the linking process.

If the Virtual I/O Server was presenting file-backed devices to the virtual I/O clients, these volume group structures need to be restored by using the **`restorevgstruct`** command. These files are in the `/home/ios/vgbackups` directory if you performed a full backup by using the **`backupios`** command.

After you restore all of the logical volume structures, the only remaining step is to restore the virtual devices that link the physical backing device to the virtual. To restore these devices, use the **`viosbr`** outputs recorded from the backup steps in "Backing up device mappings" on page 36. It is important to remember the slot numbers and backing devices when you restore these mappings.

3.1.5 NIM server resilience

It is instrumental to have a working NIM server during disaster recovery procedures. In an ideal setup, the NIM server that is responsible for restoring the Virtual I/O Server, sits independently of the virtualized environment. But, there are other solutions which can help NIM availability. Figure 3-1 details the best practices in providing a reliable NIM solution.

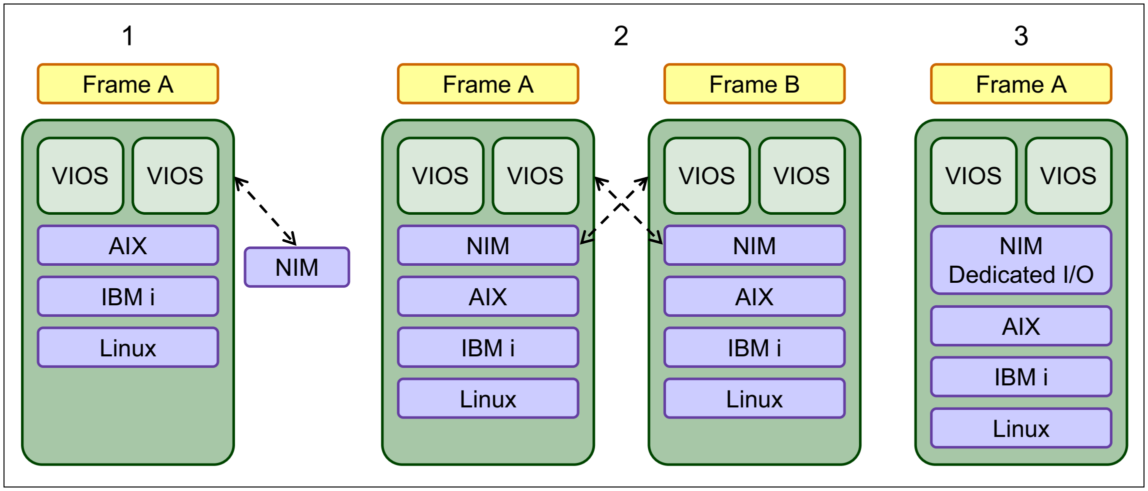


Figure 3-1 NIM resilience solutions

The following guide explains the numbers that are shown at the top of Figure 3-1 on page 41.

1. The NIM is a stand-alone server, independent of the virtualized environment.
2. The NIM that is responsible for managing the logical partitions on Frame A, is a logical partition on Frame B, and vice versa.
3. The NIM uses dedicated resources (that is, the network and storage are not managed by the Virtual I/O Server).

In all three scenarios, the NIM continues to function if one or all of the Virtual I/O Servers on a single frame are unavailable.

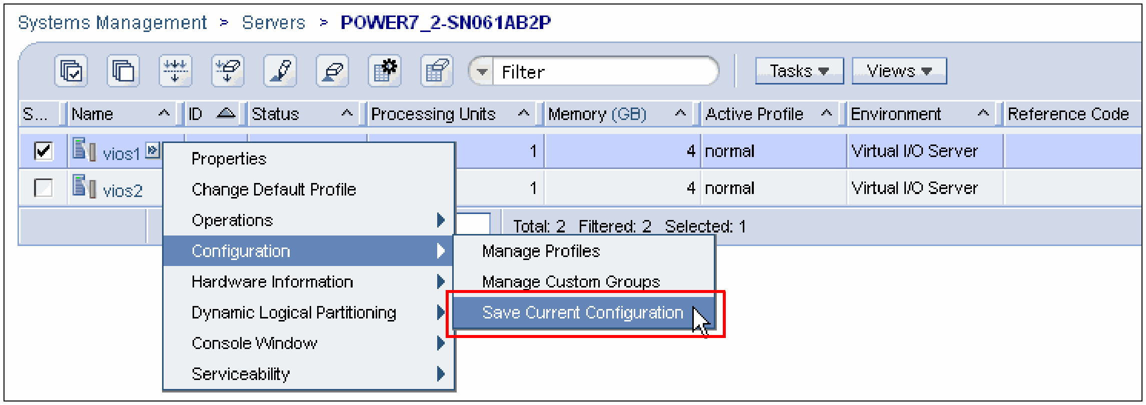
3.2 Dynamic logical partition operations

When you remove, change, or add: I/O adapters, memory, processor, or virtual adapters by using a dynamic logical partition (LPAR), they are not automatically reflected in the permanent profile of the partition. Therefore, whenever you reconfigure, we suggest that you update the profile. A good practice is to first put the changes into the partition profile. Then, make the changes dynamically to prevent them from being lost in the case of a deactivation or shut down. The changes that are made to the partition in the partition profile are not reflected in the LPAR if you perform a reboot. Changes are only reflected when you start the LPAR from a Not Activated state.

An alternative way to save the active partition assignments is by using the option **Configuration** → **Save Current Configuration**, which is selected from the partition context menu, as shown in Figure 3-2 on page 43. When you save the new profile, select a meaningful name such as `normal_new_<today_date>`, as shown in Figure 3-3 on page 43. After you test the new partition profile, rename the old partition profile to, for example, `normal_old_<today_date>` and rename or copy the new partition profile `normal_new_<today_date>` to `normal`. Remember to document what you changed in the new partition profile. In addition, you might want to clean up any old unused partition profiles from the last save or rename them now.

An alternative method is to use the Overwrite existing profile option, as shown in Figure 3-4 on page 44. Use caution when you use this option so you do not overwrite a profile which is still required.

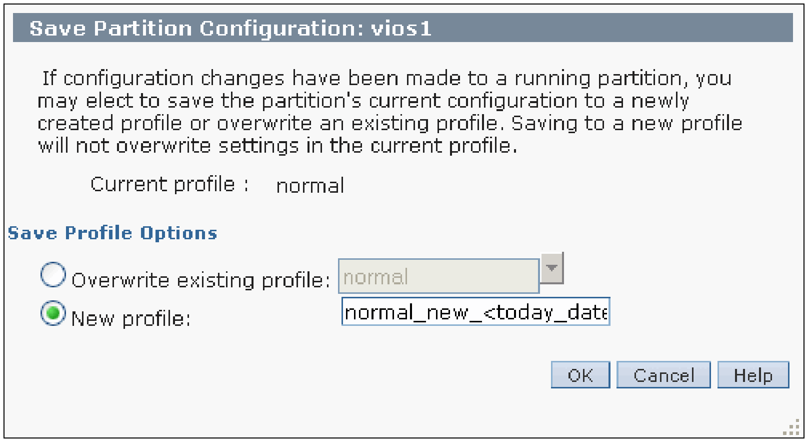
Figure 3-2 on page 43 shows the partition context menu. Example 3-3 on page 36 shows how to save the running partition profile to a new profile.


Systems Management > Servers > POWER7_2-SN061AB2P

S...	Name	ID	Status	Processing Units	Memory (GB)	Active Profile	Environment	Reference Code
☑	vios1			1	4	normal	Virtual I/O Server	
□	vios2			1	4	normal	Virtual I/O Server	

Total: 2 Filtered: 2 Selected: 1

Figure 3-2 Partition context menu


Save Partition Configuration: vios1

If configuration changes have been made to a running partition, you may elect to save the partition's current configuration to a newly created profile or overwrite an existing profile. Saving to a new profile will not overwrite settings in the current profile.

Current profile : normal

Save Profile Options

○ Overwrite existing profile: normal

● New profile: normal_new_<today_date

Figure 3-3 Save the running partition profile to a new profile

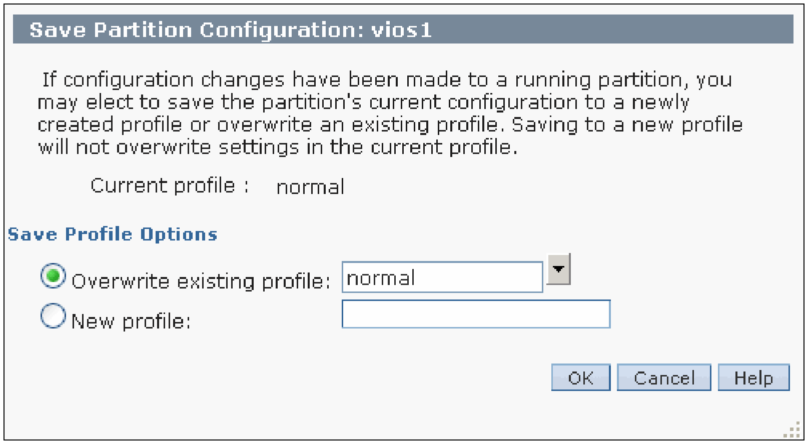
If you have multiple workloads that require multiple partition profiles, do not clean up or rename the partition profiles. Activate the profiles and do not use a date naming scheme, but one that is meaningful to this workload, such as `DB2_high`.

3.2.1 Dynamically adding virtual Fibre Channel adapters

One method of adding a virtual Fibre Channel adapter to a virtual I/O client, is by using a dynamic LPAR operation, and then modifying the permanent partition profile. This method results in a different pair of worldwide port names (WWPNs) between the active and saved partition profiles.

When a virtual Fibre Channel adapter is created for a virtual I/O client, a pair of unique WWPNs is assigned to this adapter by the Power Hypervisor. An attempt to add the same adapter at a later stage results in the creation of another pair of unique WWPNs.

When you add virtual Fibre Channel adapters into a virtual I/O client with a dynamic LPAR operation, use the *Overwrite existing profile* option to save the permanent partition profile. This option is shown in Figure 3-4. This choice results in the same pair of WWPNs in both the active and saved partition profiles.



Save Partition Configuration: vios1

If configuration changes have been made to a running partition, you may elect to save the partition's current configuration to a newly created profile or overwrite an existing profile. Saving to a new profile will not overwrite settings in the current profile.

Current profile : normal

Save Profile Options

● Overwrite existing profile: normal

○ New profile:

Figure 3-4 Overwrite the existing partition profile

3.3 Virtual media repository

The *virtual media repository,* also known as the *virtual media library*, allows the loading of disc images in the ISO format onto the Virtual I/O Server to share with virtual I/O clients. Configuration of the virtual media repository is simple, and is covered in *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590.

One of the key benefits of the virtual media repository is the ability to install operating systems onto virtual I/O clients. This feature is essentially a similar functionality to what the NIM provides, but can be much quicker. A good practice is to keep copies of the operating system gold images that can be deployed to multiple logical partitions dynamically if the NIM is unavailable.

AIX tip: On AIX, the `mkdvd` and `mkcd` commands allow for the conversion of a mksysb image to an ISO format, which can be loaded into the virtual media repository. Images in the virtual media repository, which have read-only permissions, can be shared between multiple logical partitions.

IBM i tip: On IBM i, you can use the virtual media repository as an alternate restart device for Licensed Internal Code installation.

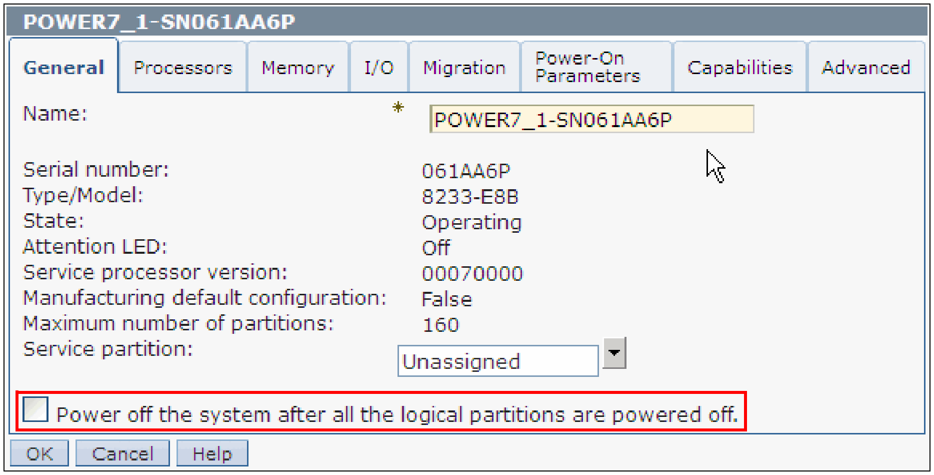
3.4 Power Systems server shutdown and startup

The following section describes best practices to shut down and start a whole IBM Power Systems server by using the HMC.

There are two types of a Power Systems server shutdown, planned and unplanned. In either case, with an emphasis on unplanned shutdowns, we suggest that you have partition startup under manual control. Power Systems server shutdowns are a rare operation, and occur mostly as a result of a disaster.

Before a planned server shutdown, save your current profiles, check if you have the correct profiles that are set as the default, and perform any other administration tasks. However, be prepared for unplanned server shutdowns and do backups regularly. For more information about backups, see 3.1, “Backing up and restoring the Virtual I/O Server” on page 32.

A best practice is to disable *Power off the system after all the logical partitions are powered off*. Figure 3-5 shows how to enable and disable this option.



Name: * POWER7_1-SN061AA6P

Serial number: 061AA6P
Type/Model: 8233-E8B
State: Operating
Attention LED: Off
Service processor version: 00070000
Manufacturing default configuration: False
Maximum number of partitions: 160
Service partition: Unassigned

☐ Power off the system after all the logical partitions are powered off.

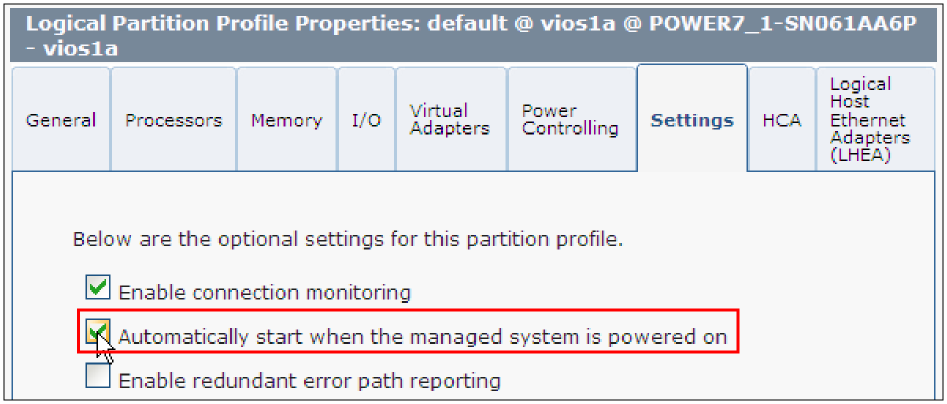
Figure 3-5 The `Power off system` option is turned off

If you use Live Partition Mobility (LPM) to optimize your power consumption, set **partition auto startup** for the Virtual I/O Servers. This setting ensures that after the server is started, it will be ready for LPM as soon as possible. For more information, see 7.3, “Live Partition Mobility” on page 109.

> **Schedule startup and shutdown:** You can schedule the Power Systems server startup and shutdown via the HMC:
> **Servers -> ServerName -> Operations -> Schedule Operations**.

Select *Automatically start when the managed system is powered on* for the Virtual I/O Servers, and for any logical partitions that do not depend on the Virtual I/O Servers. We also suggest enabling this function on logical partitions that you want to achieve the best memory and processor affinity.

To set this option in the HMC, select the logical partition and the profile that you want to start automatically. Choose **Settings** and select the check box for an automatic start, as shown in Figure 3-6.


Logical Partition Profile Properties: default @ vios1a @ POWER7_1-SN061AA6P - vios1a

Below are the optional settings for this partition profile.

☑ Enable connection monitoring

☑ Automatically start when the managed system is powered on

□ Enable redundant error path reporting

Figure 3-6 Automatically start when the managed system is powered on

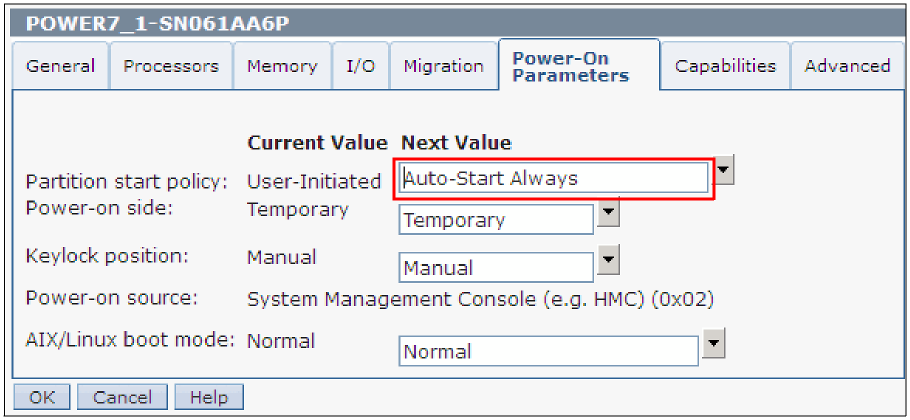
You can specify an order in which LPARs are started automatically. The automatic startup of all the logical partitions might work in your environment; however, there are some things to consider:

- After the Power Systems server starts up, for every LPAR that is being activated, new memory and processor affinity will be defined.

- If you start all the partitions at the same time, the client logical partitions with the boot devices that are dependent on the Virtual I/O Server (virtual SCSI or NPIV), wait and try again to boot. The partitions try again to boot until at least one Virtual I/O Server finishes its activation. In dual Virtual I/O Server setups, the following steps occur:
  - Your logical partition will boot after the first Virtual I/O Server comes up. This action might result in stale physical partitions (PPs) in volume groups that are mirrored through both Virtual I/O Servers.
  - Some storage paths might be in an inactive state.
  - All logical partitions might be using only a single Virtual I/O Server for network bridging.
- There might be application dependencies, for example, a Domain Name System (DNS), a Lightweight Directory Access Protocol (LDAP), a Network File System (NFS), or a database that is running on the same Power Systems server. If there are dependencies, check that these partitions activate correctly.

Consider the following options when you are defining your startup order list:

- Start the Virtual I/O Servers first or set them to start automatically.
- Set Partition start policy to **Auto-Start Always** or **Auto-Start for Auto-Recovery**, as shown in Figure 3-7. In the HMC, choose **Server -> ServerName -> Properties, Power-On Parameters**.



	Current Value	Next Value
Partition start policy:	User-Initiated	Auto-Start Always
Power-on side:	Temporary	Temporary
Keylock position:	Manual	Manual
Power-on source:	System Management Console (e.g. HMC) (0x02)	
AIX/Linux boot mode:	Normal	Normal

Figure 3-7 Partition start policy

- The LPARs with the highest memory and the highest number of processors are started first to achieve the best processor and memory affinity. For critical production systems, you might want to activate them before the Virtual I/O Servers.

- Start the LPARs that provide services for other partitions. For example, first start with the DNS and LDAP servers. Also, start the NFS server first, especially if there are NFS clients with NFS hard mounts on startup.
- Prepare the HMC commands to simplify startup. A spreadsheet might help you create those commands. To start an LPAR and open its virtual terminal on a specific system, use the commands shown in Example 3-6.

Example 3-6 HMC CLI startup commands

```
chsysstate -m <servername> -o on -r lpar -n vios1a -f <profilename>
mkvterm -m <servername> -p vios1a
```

The startup sequence does not need to be serial; many LPARs can be started at the same time. The following scenario provides a server startup example:

- Activate the critical systems with a high amount of memory and processors, first.
- Start all Virtual I/O Servers at the same time.
- After the Virtual I/O Servers are up and running, activate all your DNS, LDAP, or NFS servers at the same time.
- Activate all database LPARs at the same time. There might be a cluster that needs to be started.
- After the database LPARs are up and running, activate all the application LPARs at the same time.

Grouping partitions: You can use system profiles to group the partitions. For more information, see this website:

http://www.ibm.com/developerworks/aix/library/au-systemprofiles/

4

Networking

Virtual network infrastructure that is used in IBM Power Systems is extremely flexible and allows for a number of configuration options. This chapter describes the most common networking setups that use a dual Virtual I/O Server configuration as a best practice.

© Copyright IBM Corp. 2012. All rights reserved.

4.1 General networking considerations

Chapter 4 describes several best practice scenarios. The basic difference is determining how many Virtual I/O Servers that you use and whether you are going to use virtual local area networks (VLANs). Table 4-1 shows the most important terminology that is used in Chapter 4.

Table 4-1 Terminology that is used in this chapter

Terminology	Definition in this chapter
Link Aggregation (LA)	LA aggregates more physical connections that operate as one link. It increases network throughput and high availability. It has two basic modes of operation: Etherchannel and 802.3ad, also known as *port trunking*.
Network Interface Backup (NIB)	NIB provides a highly available connection by having a backup adapter that, by default, is not active. It is usually used on different LAN cards that are connected to different LAN switches.
VLAN tagging	Standard 802.1Q allows you to have more VLANs defined on one physical connection or a Link Aggregation. Also known as VLAN trunking or simply, *trunking*.
Virtual network infrastructure	Virtual network infrastructure implementation that is used on IBM Power Systems. The Power Hypervisor supports 802.1Q standard.
External network infrastructure	A physical network infrastructure outside of the IBM Power Systems. Although it is referred to as a *physical network infrastructure*, virtualization techniques do exist there, too.

The following contains a list of best practices for your virtual environment:

- Before you start planning your virtual network infrastructure, speak with your network administrators and specialists to synchronize your terminology.
- Keep things simple. Document them as you go.
- Use VLAN tagging. Use the same VLAN IDs in the virtual environment as they exist in the physical networking environment.
- For virtual adapters on client logical partitions (LPARs), use Port VLAN IDs. For simplification of the installation, do not configure multiple VLANs on one adapter and do not use AIX VLAN tagged interfaces on it.
- Use hot pluggable network adapters for the Virtual I/O Server instead of the built-in integrated network adapters. They are easier to service.
- Use two Virtual I/O Servers to allow concurrent online software updates to the Virtual I/O Server.

- Spread physical Virtual I/O Server resources across multiple enclosures and I/O drawers.
- Configure an IP address on a Shared Ethernet Adapter (SEA) to allow the ping feature of a SEA failover.
- Use separate virtual Ethernet adapters for Virtual I/O Server management rather than putting it on the SEA. You can lose network connectivity if you are changing SEA settings.
- Use a *vterm* when you configure the network on a Virtual I/O Server.
- Find a balance between the number of virtual Ethernet adapters and the number of virtual local area networks (VLANs) for each virtual Ethernet adapter in a SEA. Try to group common VLANs on the same virtual Ethernet adapter. For example, group production VLANs on one adapter and non-production VLANs on another.
- Use multiple virtual switches when you work in multi-tenant environments.

> **VLAN ID 3358:** Avoid the use of VLAN ID 3358 in your infrastructure. It is reserved to enable the `tme` attribute for the virtual Fibre Channel and SAS adapters.

Remember the following technical constraints during your configuration:

- The maximum number of VLANs per virtual adapter is 21 (20 VLAN IDs (VIDs) and 1 Port VLAN ID).
- The maximum number of virtual adapters for each SEA is 16.
- The maximum number of physical Ethernet adapters in a link aggregated adapter is eight for primary and one for backup.
- The maximum virtual Ethernet frame size on the Power Systems server is 65,408 bytes.

4.1.1 Shared Ethernet Adapter considerations

The following considerations pertain to Shared Ethernet Adapters (SEAs).

VLAN bridging

If you use the suggested VLAN tagging option for the virtual network infrastructure, there are three ways to define VLANs for one SEA:

- All VLANs are on one virtual adapter, one adapter under a SEA.
- One VLAN per one virtual adapter, many adapters under a SEA.
- A combination of the two configurations.

Changing a list of virtual adapters that are bridged by a SEA can be done dynamically. One SEA can have a maximum of 16 virtual Ethernet adapters. If you need to bridge more than 16 VLANs for each SEA, define more VLANs for each virtual Ethernet adapter.

To change a set of VLANs on a virtual Ethernet adapter, check whether your environment supports dynamic change. For more information, see the following website:

http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hb1/iphb1_vios_managing_vlans.htm

If your environment does not support dynamic changes, remove the adapter from the SEA and then dynamically remove the adapter from the LPAR. Alter the VLAN definitions on the virtual adapter, add it dynamically back to the system, and alter the SEA configuration. During this operation, all access to the VLANs defined on the virtual adapter is unavailable by that Virtual I/O Server. However, all client LPARs automatically fail over and use the other Virtual I/O Server. Remember to apply the changes to the partition profile.

Shared Ethernet Adapter threading

There are two Shared Ethernet Adapter (SEA) threading modes of operation: threaded and non-threaded. At the time of this writing, threaded is the default option. If your Virtual I/O Server is virtualizing both the SCSI and the network and you experience disk performance issues, consider switching the threading off. For more information, see this website:

http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/perf.html

4.1.2 Maximum transmission unit best practices

Table 4-2 on page 52 shows the maximum transmission unit (MTU) values in different environments. We suggest checking with your network administrator which MTU value is appropriate for your network. Usually you can use a higher MTU for communication within a network. When the communication must pass through a router, the MTU is 1500, by default.

Table 4-2 Typical maximum transmission units (MTUs)

Network type	MTU (bytes)
Standard maximum MTU	65,535
IBM Power Systems maximum MTU	65,390
Ethernet (gigabit)	9000

Network type	MTU (bytes)
Ethernet (10 or 100 MBps)	1500
Standard minimum MTU	68

When data over 1500 bytes per packet are sent over a network, consider switching to a larger MTU value. For example, the size of a DNS query is small, and a large MTU value has no effect here. However, backups over the network are very demanding on the bandwidth, and might benefit from a larger MTU value.

Ensure that path MTU discovery is on. Use the **pmtu** command on AIX, the **CFGTCP** command on IBM i, and the **tracepath** command on Linux.

To change the MTU on a specific adapter, follow these steps:

- On AIX, enter **chdev -a mtu=***<new mtu>* **-l en***X*
- On Linux, see the distribution documentation. The generic rule is to use the **ifconfig** command. Also, use the appropriate parameters under /proc/sys/net/
- On IBM i, use the following procedure:

  a. In the command line, type **CFGTCP** and press **Enter**.
  b. Select **option 3** (Change TCP/IP attributes).
  c. Type the command **2** (Change) next to the route you want to change.
  d. Change the MTU size in the Maximum transmission unit field. Press **Enter** to apply the change.

If you send data through a network and the data is larger than the MTU of the network, it becomes fragmented, which has a negative affect on performance. If you experience network issues after an MTU change, check the MTU for a specific remote host or network by using **ping -s** *<size>* on AIX or Linux. Or on IBM i, use **PING RMTSYS('***<remote address>***') PKTLEN(***<size>***)**.

4.1.3 Network bandwidth tuning

To achieve the maximum out of high speed adapters, consider the following best practices:

- Use the largest possible MTU that is available in your network environment.
- IBM Power Systems servers support an MTU up to 65,390. This size is sensible for huge network transfers that do not leave the Power Systems server. An example is a backup within a Power Systems server.
- Use path MTU discovery on client partitions.

- It is important to set flow control on both the switch and the Virtual I/O Server to avoid packet resending in case the receiver is busy.
- Unless you have IBM i or Linux in your environment, enable the `large_send` and `large_recieve` attributes.

For optimum performance, especially on a 10-Gb network, tuning needs to be performed. The most important settings are provided in the following list:

- Physical adapters on the Virtual I/O Server
  `jumbo_frames=yes, large_send=yes, large_receive=yes, flow_ctrl=yes`
- Shared Ethernet adapters on the Virtual I/O Server
  `jumbo_frames=yes, largesend=yes, large_receive=yes`
- Virtual ethernet adapters on the Virtual I/O Server under SEA

  `dcbflush_local=yes`
- Client partitions
  - Running AIX
    - NIB on client, if applicable
      `jumbo_frames=yes`
    - Network interface (en*X*)
      `mtu` to the largest possible value for your network, `mtu_bypass=yes`
    - Network options
      `tcp_pmtu_discover=1, udp_pmtu_discover=1`
    - To specify the MTU for a specific host or network, add a static route with the -mtu parameter. Put it in /etc/rc.net:

      ```
      route add <destination> <GW> -mtu <MTU>
      ```
  - Running IBM i, set the MTU as described in 4.1.2, “Maximum transmission unit best practices” on page 52
  - Running Linux, see the documentation for your distribution. The generic rule is to use the **`ifconfig`** command. Also, set the appropriate parameters under `/proc/sys/net/`

4.2 Single Virtual I/O Server

The Virtual I/O Server is reliable and stable. If you are using a single Virtual I/O Server, the best practices for networking are simple: Keep every connection highly available. However, for concurrent updates and maintenance, you must either have two Virtual I/O Servers or use the Live Partition Mobility (LPM):

- Use LA wherever possible. You achieve high availability and an increase in network performance.
- If you cannot use LA, use NIB.
- Single Virtual I/O Server setups are common on hardware that have a limited number of ports. If possible, use VLAN tagging.
- Use ports from different adapters that are in different enclosures and I/O drawers.

4.3 Virtual network redundancy and failover technology

There are multiple ways to provide virtual network redundancy and failover capability. Broadly, they can be classified into two types that depend on which entity makes the failover decision:

- A server-side failover solution that is based on the SEA failover capability and configuration (in the Virtual I/O Server).
- A client-side failover solution that is based on the EtherChannel NIB failover capability and configuration (in the client LPAR).

The main advantage of a server-side failover solution is that it simplifies the client configuration because you only need to create a single virtual Ethernet adapter in the client, which is served by two SEAs that are configured in the high availability mode (failover configuration) in two Virtual I/O Servers. While the client-side failover solution requires two virtual Ethernet adapters and an additional NIB configuration for each client LPAR.

Furthermore, for the virtual Ethernet, since the link status is always *UP*, the NIB failover must rely on the ping address feature. Also, there are limitations in the case of a failback. See 4.3.1, “Dual Virtual I/O Server with VLAN tagging” on page 56.

One clear advantage that the client-side failover solution had over the server-side failover solution was that the client side option allows load sharing. With the new load sharing feature that is added to the SEA high availability configuration, this advantage is reduced.

4.3.1 Dual Virtual I/O Server with VLAN tagging

It is essential that you balance the network workload through as many adapters as possible. Especially in a 10-Gb network environment, where adapters can be costly, it is important to use your entire network infrastructure. The following configurations are two of the best options to use your network adapters in a dual Virtual I/O Server environment:

- The use of two virtual switches, one virtual switch for each Virtual I/O Server. With this option, you can balance network bridging between Virtual I/O Servers for every client and every VLAN, individually.
- The use of SEA failover with load sharing. With this option, you can balance network bridging between Virtual I/O Servers for every VLAN, individually. All LPARs with a specific VLAN are bridged by the same Virtual I/O Server. VLAN distribution between Virtual I/O Servers is automatic.

IBM i: For IBM i, we suggest that you use a SEA failover with load sharing. If you use a virtual Interface in a dual virtual switch configuration and there is a network path failure, IBM i does not detect the loss of connectivity through the virtual Ethernet adapter.

Dual virtual switch configuration

Figure 4-1 on page 57 shows two Virtual I/O Servers and two LPARs with a dual virtual switch configuration. Each LPAR has a different set of VLANs.

Each Virtual I/O Server has a Link Aggregation of two physical 1-Gb ports. All connections operate in trunking mode and are used to trunk VLAN IDs 1,2, and 3. Those VLAN IDs are used in the external network infrastructure (physical switches), as well. Virtual Ethernet adapters that are in Virtual I/O Servers operate in the trunking mode. Therefore, the VLAN tag is not removed from the Ethernet packet when a packet is received from the external network infrastructure.

In the example that is shown in Figure 4-1 on page 57, the shared Ethernet adapter connects to two virtual Ethernet adapters. There is one adapter for VLAN IDs 1,2, and one adapter for VLAN IDs 3,4,5. You can define up to 20 VLANs for each virtual Ethernet adapter in 802.1Q mode.

All the virtual Ethernet adapters of each Virtual I/O Server connect to one virtual switch. Try to avoid defining virtual Ethernet adapters from different Virtual I/O Servers in one virtual switch. This way, you eliminate the chance of networking issues because of a network misconfiguration.

Port VLAN ID: The PVID that is used on the virtual Ethernet adapter, which makes up the SEA, can be a PVID which is not used in your network.

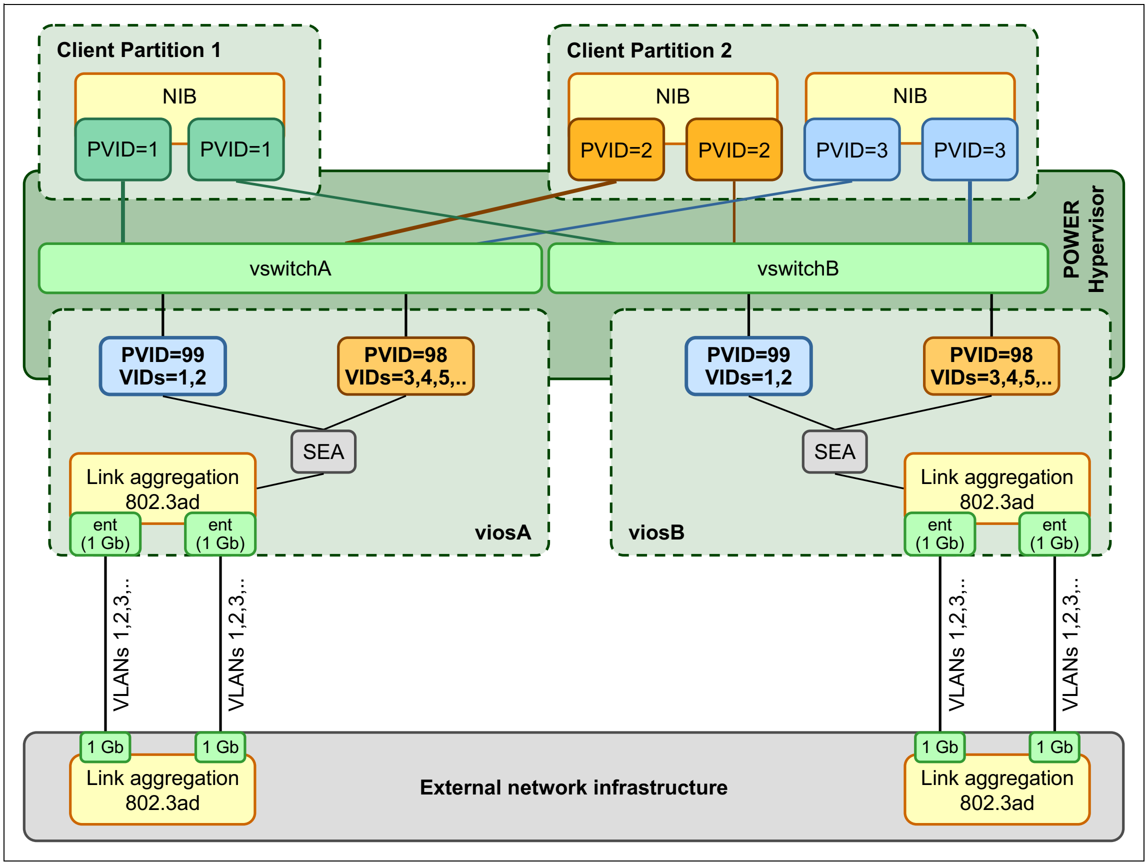


Figure 4-1 Dual Virtual I/O Server configuration with two virtual switches

Each client partition has a NIB that is made up of two virtual Ethernet adapters for every VLAN to which it connects. For details about a NIB setup that uses virtual Ethernet adapters, see 4.1.1, “Shared Ethernet Adapter considerations” on page 51. For more information and to obtain a setup guide, see this website:

http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101752

Network Interface Backup on client LPARs

If you use the Network Interface Backup (NIB) with virtual Ethernet adapters on AIX, the NIB is unable to obtain a link status because the virtual Ethernet adapter is always *up*. This configuration has a few impacts:

- The NIB must ping an IP address that is outside of the Power Systems server. To avoid the error log from filling up with NIB entries, apply an error log filter for error ID `5561971C` by using the **`errupdate`** command.
- The error log contains entries about the NIB being unable to perform certain operations. Best practice is to ping a gateway of a specific network, because this IP address is always up. If this IP is not reachable, something is misconfigured on either the Virtual I/O Server or the external network infrastructure.
- In a failover scenario, the NIB does not fail back to the primary adapter when the network path through the primary adapter recovers. It fails back only if the network path through the backup adapter fails.

If the address that is configured to ping in a NIB does not respond, the error log fills up rapidly with errors about the NIB failover and that the destination is unreachable. Message IDs `6169289D` and `DC32143C` are logged to the error log. We suggest to not filter these errors, but check and clean them periodically until the problem is resolved. Also, consider expanding the AIX error log size by using the **`/usr/lib/errdemon -s`** *<size>* command.

Linux and IBM i tips: For Linux systems, check the distribution documentation about bonding. For IBM i, we do not suggest that you use this functionality on virtual Ethernet adapters.

4.3.2 Shared Ethernet Adapter failover with load sharing

Another approach is to use the traditional Shared Ethernet Adapter (SEA) failover setup, enhanced with load sharing, as shown in Figure 4-2 on page 59. Load balancing between SEAs is on a per virtual Ethernet adapter basis. This means that all VLANs defined in one virtual Ethernet adapter that are used by a SEA, are bridged by one Virtual I/O Server.

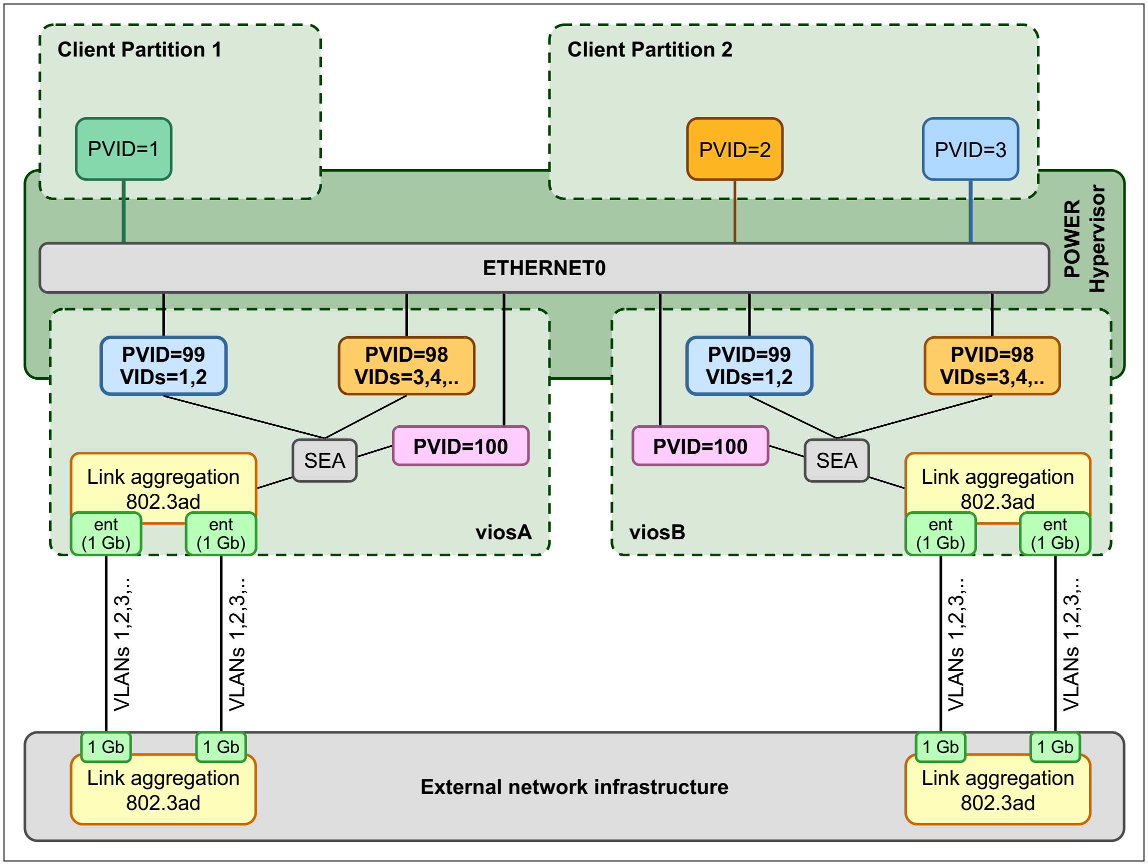


Figure 4-2 Dual Virtual I/O Server configuration with SEA and load balancing

The SEA is connected to two virtual Ethernet adapters; each adapter has a different set of VLAN IDs. Virtual Ethernet adapters on different Virtual I/O Servers must have the same set of VLANs and different trunk priorities.

In addition, the SEA is connected to another virtual Ethernet adapter with a PVID=100 that is used as a SEA control channel. The control channel is used for SEA heartbeating and exchanging information between the two SEA adapters on the set of VLAN IDs that each SEA bridges.

For more information and to obtain a setup guide, see this website:

http://www.ibm.com/support/docview.wss?uid=isg3T7000527

5

Storage

The Virtual I/O Server has different means to provide storage access to virtual I/O clients. This chapter addresses the best practices on the different methods for presenting and managing storage on the Virtual I/O Server and virtual I/O clients.

© Copyright IBM Corp. 2012. All rights reserved.

5.1 Storage Considerations

This section describes the various storage requirements from the Virtual I/O Server perspective.

We also describe the different methods of providing storage access to a virtual I/O client:

Virtual Small Computer System Interface

The virtual Small Computer System Interface (SCSI) was the first method for IBM PowerVM to provide virtual storage to virtual I/O clients. With the SCSI, storage resources that are presented to the virtual clients, need to be mapped to physical storage by the Virtual I/O Server. The most current versions of the Virtual I/O Server introduced new features, such as file-backed devices, virtual optical devices, and virtual tape devices.

Virtual Fibre Channel

Virtual Fibre Channel is the PowerVM implementation of the Fibre Channel industry standard *N_Port ID Virtualization* (NPIV).

NPIV can share a physical Fibre Channel adapter across multiple virtual I/O clients. It has the advantage to provide direct Fibre Channel access to the virtual clients from the external storage devices, such as disks arrays and tape libraries. It does not require an extensive storage mapping effort from the Virtual I/O Server.

For more information about setting up and managing virtual storage, see *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590**.**

5.1.1 Virtual I/O Server rootvg storage

There are many aspects to consider regarding the storage management at the Virtual I/O Server. One of most common questions is whether to install the *root volume group* (rootvg) of the Virtual I/O Server on the internal or external storage. There are some advantages to both options. The choice for a booting device depends on your infrastructure and on the number of internal disks, disk adapters, and Fibre Channel adapters that are available.

Newer server technology, such as IBM PureSystems, also adds to the decision process because their high density design favors the use of SAN booting because of the number of internal drives and adapters that are available.

The best practice for booting a Virtual I/O Server is to use internal disks rather than external SAN storage, if the server architecture and available hardware allows. The following list provides reasons for booting from internal disks:

- The Virtual I/O Server does not require specific multipathing software to support the internal booting disks. This configuration helps when you perform maintenance, migration, and update tasks.
- The Virtual I/O Server does not need to share Fibre Channel adapters with virtual I/O clients, which helps if a Fibre Channel adapter replacement is required.
- The virtual I/O clients might have issues with the virtual SCSI disks presented by the Virtual I/O Server that is backed by SAN storage. If so, the troubleshooting can be performed from the Virtual I/O Server.
- We do not recommend the allocation of logical volume backing devices for virtual I/O clients in the rootvg of the Virtual I/O Server.
- The SAN design and management is less critical.
- Access to the dump device is simplified.

A SAN boot provides the following advantages:

- SAN hardware can accelerate booting through their cache subsystems.
- Redundant Array of Independent Disks (RAID) and mirroring options might be improved.
- A SAN boot is able to take advantage of advanced features that are available with SAN storage.
- Easy to increase capacity.
- Provides smaller server hardware footprints.
- Zones and cables can be set up and tested before client deployment.

In general, avoid single points of failure. Following is a list of best practices for maintaining the Virtual I/O Server availability:

- Mirror rootvg by using the **`mirrorios -defer hdiskX`** command, and then reboot the Virtual I/O Server at your earliest convenience.
- If rootvg is mirrored, use disks from two different disk controllers. Ideally, those disk controllers are on different Peripheral Component Interconnect (PCI) busses.

> **Booting from external storage:** If you are booting from external storage, use multipathing software, with at least two Fibre Channel adapters that are connected to different switches.

For more information about installing the Virtual I/O Server on a SAN, see:

- *IBM Flex System p260 and p460 Planning and Implementation Guide*, SG24-7989

  http://www.redbooks.ibm.com/redbooks/pdfs/sg247989.pdf
- *IBM System Storage DS8000 Host Attachment and Interoperability*, SG24-8887

  http://www.redbooks.ibm.com/redbooks/pdfs/sg248887.pdf

5.1.2 Multipathing

Multipathing is a best practice in terms of performance and redundancy; it can be implemented from both the virtual I/O client and the Virtual I/O Server. Multipathing provides load balancing and failover capabilities if an adapter becomes unavailable. This process also helps if there is a SAN problem on the path to the external storage.

A virtual SCSI configuration requires more management from the Virtual I/O Server than NPIV. Figure 5-1 on page 65 shows an example of dual Virtual I/O Servers that provide a multipath SAN disk to a virtual I/O client. In this case, multipathing is managed from both the virtual I/O client and Virtual I/O Server.

Multipathing for clients with dual Virtual I/O Servers

Whether you implement multipathing at the Virtual I/O Server, the virtual I/O client, or both, it is a best practice to have dual Virtual I/O Servers that are backed by external storage. The following list provides reasons to back the servers by external storage:

- Unscheduled outages, such as physical device failures and human errors.
- Scheduled operations that are required for Virtual I/O Server maintenance.
- High availability and load balancing of storage I/Os.
- Automatic path failover protection.
- Prevention of single points of failure on the path to the external storage.

With NPIV, multipathing is only managed from the virtual I/O client. For more information, see section 5.4, “N-Port ID Virtualization” on page 88.

Figure 5-1 on page 65 shows a virtual SCSI client with multipathing from dual Virtual I/O Servers.

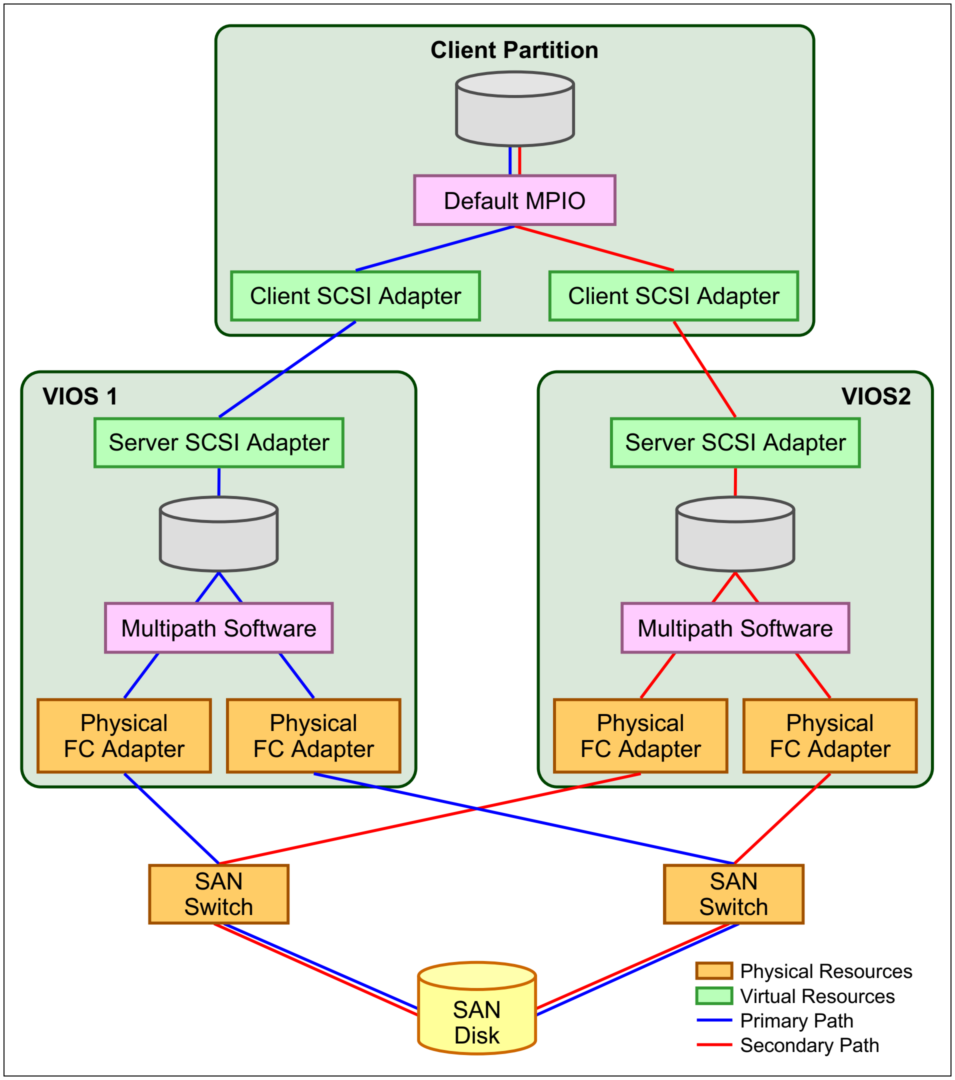


Figure 5-1 Virtual SCSI client with multipathing from dual Virtual I/O Servers

For more information about virtual I/O client configuration, see section 5.2.2, “Configuring the Virtual I/O Server with a virtual SCSI” on page 71.

Supported storage on the Virtual I/O Server

The Virtual I/O Server supports a number of storage devices. Contact your storage vendor for more information about Virtual I/O Server support. A list of the current supported storage solutions is available at this website:

http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html#solutions

Confirm that the Virtual I/O Server supports the multipathing device driver that you plan to use. For example, the Subsystem Device Driver Path Control Module (SDDPCM), is the path control module that is designed to support a multipathing configuration with most of the IBM storage servers. At the time of this writing, the SDDPCM does not support the Virtual I/O Server with the IBM DS4000®, DS5000, or DS3950 storage subsystems. If you need SDDPCM support for the IBM DS8000® or SVC, you must use the **manage_disk_drivers** command to ensure that the driver for your DS4000 and DS5000 is set to *AIX_APPCM*, not to *AIX_SDDAPPCM*. Example 5-1 shows this command. This setting is valid for both the virtual SCSI clients and Virtual I/O Server usage.

For more information about the SDDPCM, see this website:

http://www.ibm.com/support/docview.wss?uid=ssg1S4000201

Example 5-1 Manage the DS4000 and DS5000 disk driver

```
$ oem_setup_env
# manage_disk_drivers -l
Device              Present Driver        Driver Options
2810XIV             AIX_AAPCM             AIX_AAPCM,AIX_non_MPIO
DS4100              AIX_APPCM             AIX_APPCM,AIX_fcparray,AIX_SDDAPPCM
DS4200              AIX_APPCM             AIX_APPCM,AIX_fcparray,AIX_SDDAPPCM
DS4300              AIX_APPCM             AIX_APPCM,AIX_fcparray,AIX_SDDAPPCM
DS4500              AIX_APPCM             AIX_APPCM,AIX_fcparray,AIX_SDDAPPCM
DS4700              AIX_APPCM             AIX_APPCM,AIX_fcparray,AIX_SDDAPPCM
DS4800              AIX_APPCM             AIX_APPCM,AIX_fcparray,AIX_SDDAPPCM
DS3950              AIX_APPCM             AIX_APPCM,AIX_SDDAPPCM
DS5020              AIX_APPCM             AIX_APPCM,AIX_SDDAPPCM
DCS3700             AIX_APPCM             AIX_APPCM
DS5100/DS5300       AIX_APPCM             AIX_APPCM,AIX_SDDAPPCM
DS3500              AIX_APPCM             AIX_APPCM
XIVCTRL             MPIO_XIVCTRL          MPIO_XIVCTRL,nonMPIO_XIVCTRL
```

SDDPCM: If you need to have SDDPCM installed or upgraded, it toggles the DS4000 and DS5000 device driver management to SDDPCM (*AIX_SDDAPPCM*). Check and change the parameter before you reboot the Virtual I/O Server.

For more information, see the SDDPCM for AIX and Virtual I/O Server support matrix at this website:

http://www.ibm.com/support/docview.wss?rs=540&uid=ssg1S7001350#AIXSDDPCM

5.1.3 Mixing virtual SCSI and NPIV

It is possible to mix a virtual Small Computer System Interface (SCSI) and N-Port ID Virtualization (NPIV) within the same virtual I/O client. You can have rootvg or booting devices that are mapped via virtual SCSI adapters, and data volumes that are mapped via NPIV.

Mixing NPIV and a virtual SCSI has advantages and disadvantages, as shown in Table 5-1.

Table 5-1 Booting from a virtual SCSI with NPIV for data

Advantages	It makes multipathing software updates easier for data disks.
	You can use the Virtual I/O Server to perform problem determination when virtual I/O clients have booting issues.
Disadvantages	Requires extra management at the Virtual I/O Server level.
	Live Partition Mobility (LPM) is easier with NPIV.

5.1.4 Fibre Channel adapter configuration

The following recommendations apply to both a virtual SCSI and NPIV. There are two levels of configuration for the Fibre Channel adapters at the Virtual I/O Server:

fscsiX	Fibre Channel protocol driver.
fcsX	Fibre Channel adapter driver.

Fibre Channel protocol driver attributes

The *fscsi* devices include specific attributes that must be changed on Virtual I/O Servers. It is a best practice to change the `fc_err_recov` and the *dyntrk* attributes of the fscsi device. Both attributes can be changed by using the **`chdev`** command, as shown in Example 5-2 on page 68.

Example 5-2 The fscsi attribute modification

```
$ chdev -dev fscsi0 -attr fc_err_recov=fast_fail dyntrk=yes
fscsi0 changed
```

Changing the `fc_err_recov` attribute to `fast_fail` fails any I/Os immediately if the adapter detects a link event, such as a lost link between a storage device and a switch. The `fast_fail` setting is only recommended for dual Virtual I/O Server configurations. Setting the `dyntrk` attribute to `yes` allows the Virtual I/O Server to tolerate cable changes in the SAN.

It is a best practice to change the fscsi attributes before you map the external storage so that you do not need to reboot the Virtual I/O Server.

Fibre Channel device driver attributes

The *fcs* devices include specific attributes that can be changed on Virtual I/O Servers. These attributes are `num_cmd_elems` and `max_xfer_size.` As a best practice, complete a performance analysis on the adapters and change these values, if needed. This analysis can be done with tools such as the **`fcstat`** command as shown in Example 5-3 on page 69.

num_cmd_elems Modifies the number of commands that can be queued to the adapter. Increasing `num_cmd_elems` decreases the `No Command Resource Count`. Increasing `num_cmd_elems` also makes it more likely to see `No Adapter Elements Count`, or `No DMA Resource Count` increasing.

max_xfer_size Has an effect on the direct memory access (DMA) region size that is used by the adapter. Default `max_xfer_size` (0x100000) gives a small DMA region size. Tuning up `max_xfer_size` to 0x200000 provides a medium or large DMA region size, depending on the adapter.

Fscsi and fcs changes: Any change to the fscsi and fcs attributes need to be first checked with your storage vendor.

Example 5-3 on page 69 shows that there is no need to change the `max_xfer_size`, since the `No DMA Resource Count` did not increase. In the same example, consider increasing the `num_cmd_elem` since the `No Command Resource Count` increased. These values are measured since the last boot or last reset of the adapter statistics.

Example 5-3 Monitoring fcs adapter statistics

```
$ fcstat fcs0|grep -p 'Driver Information'
IP over FC Adapter Driver Information
  No DMA Resource Count: 0
  No Adapter Elements Count: 395

FC SCSI Adapter Driver Information
  No DMA Resource Count: 0
  No Adapter Elements Count: 395
  No Command Resource Count: 2415
```

Both attributes can be changed by using the **chdev** command, as shown in Example 5-4.

Example 5-4 fcs attributes modification

```
$ lsdev -dev fcs0 -attr max_xfer_size,num_cmd_elems
value

0x100000
500
$ chdev -dev fcs0 -attr num_cmd_elems=1024 -perm
fcs0 changed
$ chdev -dev fcs0 -attr max_xfer_size=0x200000 -perm
fcs0 changed
```

In summary, consider the following best practices for the Fibre Channel device driver values:

- Do not increment these values without an appropriate analysis.
- Increment these values gradually.
- Reboot the Virtual I/O Servers for these changes to take effect.

Example 5-5 demonstrates how to check whether there is a need to increase the `num_cmd_elem` by using the SDDPCM commands. We can verify that the value of 869 did not reach the `num_cmd_elems` limit of 1024.

Example 5-5 Check num_cmd_elems using the SDDPCM **pcmpath** *command*

```
$ oem_setup_env
# pcmpath query adapter

Total Dual Active and Active/Asymmetric Adapters : 6

Adpt#    Name    State     Mode             Select     Errors  Paths  Active
    0 fscsi16   NORMAL   ACTIVE               3032          0     16      16
    1 fscsi17   NORMAL   ACTIVE              2891          0     16      16
    2  fscsi0   NORMAL   ACTIVE         111857402          0    150     150
    3  fscsi4   NORMAL   ACTIVE         111833254          0    150     150
    4  fscsi8   NORMAL   ACTIVE         111664490          0    150     150
  > 5 fscsi12   NORMAL   ACTIVE         111487891          0    150     150
{kfw:root}/ # pcmpath query adaptstats 5 aa

Adapter #:  5
=============
                Total Read  Total Write  Active Read  Active Write   Maximum
I/O:             106925273     33596805            0             0       869
SECTOR:         3771066843   1107693569            0             0     27784

{kfw:root}/ # lsattr -El fcs12 -a num_cmd_elems
num_cmd_elems 1024 Maximum number of COMMANDS to queue to the adapter True
```

5.2 Virtual Small Computer System Interface

With a virtual Small Computer System Interface (SCSI), a Virtual I/O Server maps storage resources to virtual I/O clients. The Virtual I/O Server provides an abstraction layer to the virtual I/O clients. Section 5.2 covers the following virtual SCSI mapping options:

- Physical volumes
- Logical volumes
- File backed devices
- Logical units that are mapped from Shared Storage Pools (SSPs)
- Virtual optical devices
- Virtual tape devices

IBM i functionality: IBM i can host a virtual SCSI server adapter similarly to a Virtual I/O Server. It can also provide virtual disk units, virtual optical drives, and virtual Ethernet adapters to the clients. This functionality is not described in this book because it is not a PowerVM feature.

5.2.1 When to use a virtual Small Computer System Interface

A virtual SCSI is the only choice when there is no SAN infrastructure available. This interface is also the choice for sharing internal optical devices and presenting Shared Storage Pool logical units.

The best practice for selecting between these virtual SCSI options, is that the disk devices that are backed by SAN storage, are exported as physical volumes. With internal devices, if you have enough for all the virtual I/O clients, assign the entire disk as a backing device. However, if you have a limited number of disks, create storage pools so that storage can be managed from the Virtual I/O Server.

IBM i

A virtual SCSI allows for the connection to external storage, which natively does not support an IBM i block size. IBM i, for historical reasons, works with 520 bytes per sector of formatted storage. However, the Virtual I/O Server uses industry-standard 512 bytes per sector. In this case, the Hypervisor manages the block size conversion for the IBM i virtual I/O clients.

Note: With virtual SCSI, mirroring of client volume groups must be implemented at the level of the client. The logical volume backing device cannot be mirrored on the Virtual I/O Server. Ensure that the logical volume backing devices sit on different physical volumes and are served from dual Virtual I/O Servers.

5.2.2 Configuring the Virtual I/O Server with a virtual SCSI

Managing, documenting, and tracking the virtual to physical resources configuration is important, especially when it comes to large infrastructures. It is also a best practice for problem determination because it is usually not the time to determine how your virtual to physical relationship is built.

As stated in 1.3.9, “Slot numbering and naming conventions” on page 12, we want to emphasize that you need to implement a naming convention for your virtual devices. Although it must be managed for NPIV clients, it is sensitive for virtual SCSI mapped resources.

Virtual adapters slot considerations

There are many different options to establish a virtual adapter slot numbering convention. Slot numbers are shared between virtual storage and virtual network devices. In complex systems, there tend to be far more storage devices than network devices because each virtual SCSI device can communicate with only one server or client. We suggest reserving the slot numbers through 20 for network devices on all logical partitions to keep the network and storage devices grouped.

In simple environments, the management can be simplified by keeping virtual adapter slot numbers consistent between the client and the server. Start the virtual SCSI slots at 20, then add the `lpar_id` of the client to this base. Example 5-6 shows the HMC command output.

Example 5-6 HMC virtual slots listing with a single Virtual I/O Server

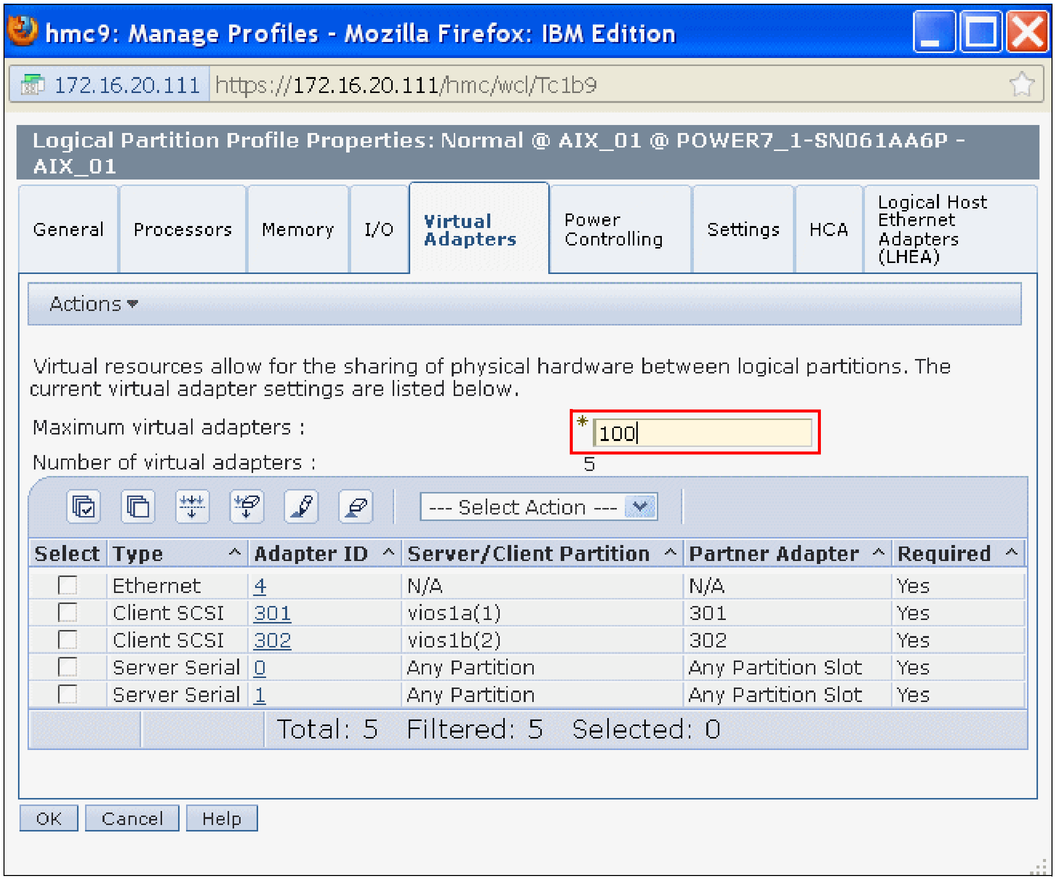
```
hscroot@hmc9:~>lshwres -r virtualio --rsubtype scsi -m
POWER7_1-SN061AA6P --level lpar -F
lpar_name,lpar_id,slot_num,adapter_type,remote_lpar_name,remote_lpar_id
,remote_slot_num | grep AIX_02

vios1a,1,25,server,AIX_02,5,25
AIX_02,5,25,client,vios1a,1,25
```

In Example 5-6, `vios1a` has slot 25, which maps to AIX_02 slot 25.

With dual Virtual I/O Servers, the adapter slot numbers can be alternated from one Virtual I/O Server to the other. The first Virtual I/O Server can use odd-numbered slots, and the second can use even-numbered slots. In a dual servers scenario, allocate slots in pairs, with each client using two adjacent slots such as `21` and `22`, or `33` and `34`.

> **Slot numbering scheme:** The slot numbering scheme can be altered when a Live Partition Mobility (LPM) migrates a logical partition.

Increase the *Maximum virtual adapters* value above the default value of 10 when you create a logical partition (LPAR) profile. The appropriate number for your environment depends on the virtual adapters slot scheme that you adopt. The allocation needs to be balanced because each unused virtual adapter slot uses a small amount of memory. Figure 5-2 on page 73 is an example of the Maximum virtual adapters value in a logical partition profile.


hmc9: Manage Profiles - Mozilla Firefox: IBM Edition
172.16.20.111
https://172.16.20.111/hmc/wcl/Tc1b9
General
Processors
Memory
I/O
Virtual Adapters
Power Controlling
Settings
HCA
Logical Host Ethernet Adapters (LHEA)
Actions
--- Select Action ---
OK
Cancel
Help

Logical Partition Profile Properties: Normal @ AIX_01 @ POWER7_1-SN061AA6P - AIX_01

Virtual resources allow for the sharing of physical hardware between logical partitions. The current virtual adapter settings are listed below.

Maximum virtual adapters : * 100

Number of virtual adapters : 5

Select	Type	Adapter ID	Server/Client Partition	Partner Adapter	Required
□	Ethernet	4	N/A	N/A	Yes
□	Client SCSI	301	vios1a(1)	301	Yes
□	Client SCSI	302	vios1b(2)	302	Yes
□	Server Serial	0	Any Partition	Any Partition Slot	Yes
□	Server Serial	1	Any Partition	Any Partition Slot	Yes

Total: 5 Filtered: 5 Selected: 0

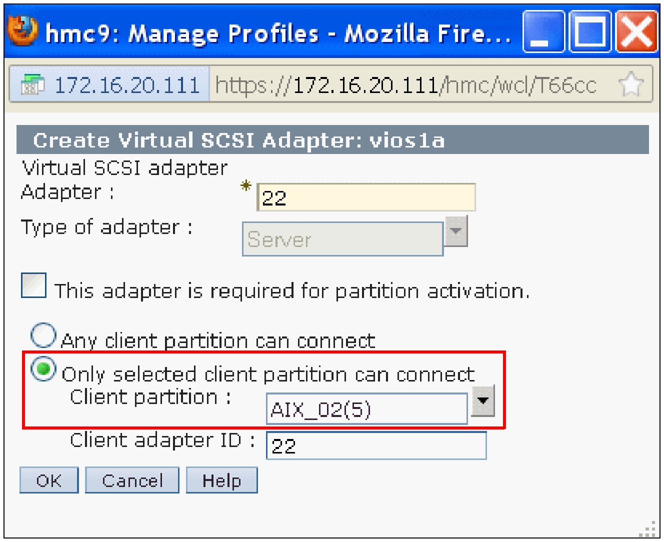
Figure 5-2 Changing the default Maximum virtual adapters number

> **Partition profile:** The maximum number of virtual adapter slots that are available on a partition is set by the profile of the partition. It is required to reactivate the LPAR to have the profile taken into account.

It is a best practice to use separate virtual adapter pairs for different types of backing devices:

- For UNIX client logical partitions, do not map boot and data disks on the same virtual adapter. As an example, for AIX, have rootvg and data volume groups on separate virtual SCSI server adapters.
- Do not map shared storage pool logical units, logical volumes, and physical volumes on the same virtual SCSI server adapter.

- For optimum performance and availability, do not share a *vhost* to map different types of physical volumes. The *max_transfer_size* for storage devices might be different. The max_transfer_size is negotiated when the SCSI adapter of the virtual client is first configured. If a new disk is mapped to the related vhost with a higher max_transfer_size, it cannot be configured at the client side. It might require a reboot.
- In a SAN environment, a LUN is assigned to the Virtual I/O Server Fibre Channel adapter. The Virtual I/O Server maps the LUN to the vhost that is associated to a virtual SCSI client adapter. From a security perspective, ensure that the specific client is connected to the virtual Fibre Channel server adapter, by explicitly selecting a partition, as shown in Figure 5-3.


hmc9: Manage Profiles - Mozilla Fire...
172.16.20.111 https://172.16.20.111/hmc/wcl/T66cc
OK
Cancel
Help

Create Virtual SCSI Adapter: vios1a

Virtual SCSI adapter

Adapter : * 22

Type of adapter : Server

□ This adapter is required for partition activation.

○ Any client partition can connect

● Only selected client partition can connect

Client partition : AIX_02(5)

Client adapter ID : 22

Figure 5-3 Select the appropriate virtual I/O client

Naming conventions

The choice of a naming convention is essential for tracking virtual to physical relationship, especially in virtual SCSI configurations.

If you want to uniquely identify the virtual client target devices, use a convention similar to the following definitions:

<client_lpar_name><bd_type><client_vg><hdisk number> :

client_lpar_name This field needs to be a representative subset, not too long.

bd_type You want to include the type of backing device used:
<L> for Logical Volume

<D> for physical disk or LUN mapping

<S> for Shared Storage Pool mapping

<V> for Virtual optical devices

<T> for Tape devices

client_vg For an AIX client, you would put a subset of the VG name.

hdisk_number You can start with hd0, then, increment. The disk number can change at the client partition. Choosing the same number is not the best idea. It is probably better to increment it.

Example 5-7 shows how you can use this naming convention.

Example 5-7 Example of a virtual storage mapping convention

```
$ mkvdev -vdev hdisk10 -vadapter vhost35 -dev Lpar1DrvgHd0
Lpar1DrvgHd0 Available
$ mkvdev -vdev hdisk11 -vadapter vhost36 -dev Lpar1DsapHd1
Lpar1DsapHd1 Available
```

Note: The **mkvdev** command currently cannot exceed 15 characters in length.

Naming convention for external physical volumes

It is a best practice to share and plan a naming convention with the SAN administration team. For example, commands, such as the AIX **mpio_get_config** command on a DS4000 and DS5000, or **svcinfo** on a SAN Volume Controller (SVC) cluster, help to identify a LUN with logical names.

Most of the storage arrays have similar commands that are meant to provide information that is relevant and useful to have an appropriate naming convention. Appropriate naming facilitates the administration and problem determination.

In Example 5-8, we take the benefit of the `user label` from a DS4800 as the virtual target device name.

Example 5-8 DS4800 **mpio_get_config** *command*

```
$ oem_setup_env
# mpio_get_config -Av
Frame id 0:
    Storage Subsystem worldwide name: 60ab800114b1c00004fad119f
    Controller count: 2
    Partition count: 1
    Partition 0:
    Storage Subsystem Name = 'DS4800POK-3'
        hdisk#           LUN #   Ownership          User Label
        hdisk6               3   A (preferred)      aix02DrvgHd0
        hdisk7               4   B (preferred)      aix02DdvgHd1
# exit
$ mkvdev -vdev hdisk6 -vadapter vhost33 -dev aix02DrvgHd0
aix02DrvgHd0 Available
$ lsmap -vadapter vhost33
SVSA            Physloc                                      Client Partition ID
--------------- -------------------------------------------- ------------------
vhost33         U8233.E8B.061AA6P-V1-C51                     0x00000005

VTD                   aix02rvgHd0

Status                Available
LUN                   0x8100000000000000
Backing device        hdisk6
Physloc               U5802.001.0086848-P1-C2-T1-W201600A0B829AC12-L3000000000000
Mirrored              false
```

oem_setup_env: Using oem_setup_env is not a best practice unless directed by IBM support, but is permitted for some storage configuration.

5.2.3 Exporting virtual Small Computer System Interface storage

This section provides best practices for exporting storage by using a virtual Small Computer System Interface (SCSI) to virtual I/O clients.

Logical volumes

The Virtual I/O Server can export logical volumes to virtual I/O clients. This method does have some advantages over physical volumes:

- Logical volumes can subdivide physical disk devices between different clients.
- System administrators with AIX experience are already familiar with the Logical Volume Manager (LVM).

In this section, the term *volume group* refers to both volume groups and storage pools. Also, the term *logical volume* refers to both logical volumes and storage pool backing devices.

Logical volumes cannot be accessed by multiple Virtual I/O Servers concurrently. Therefore, they cannot be used with Microsoft Multipath I/O (MPIO) on the virtual I/O client. Multipathing needs to be managed at the Virtual I/O Server, as explained in 5.1.2, “Multipathing” on page 64.

The following list provides best practices for logical volume mappings:

- Avoid the use of a rootvg on the Virtual I/O Server to host exported logical volumes. Certain types of software upgrades and system restores might alter the logical volume to target device mapping for logical volumes within rootvg, requiring manual intervention. Also, it is easier to manage Virtual I/O Server rootvg disk replacement when it does not have virtual I/O clients that use logical volumes as backing devices.
- The default storage pool in the Integrated Virtualization Manager (IVM) is the root volume group of the Virtual I/O Server. Ensure that you create and choose different storage pools to host client backing devices.
- Although logical volumes that span multiple physical volumes are supported, it is best if a logical volume fully resides on a single physical volume for optimum performance and for maintenance.
- Mirror the disks in the virtual I/O client. Ensure that the logical volume backing devices are on different physical disks. This configuration also helps with physical disk replacements at the Virtual I/O Server.
- In dual Virtual I/O Server configurations, if one server is rebooted, ensure that the mirroring is synced on the virtual I/O clients.

Exporting physical volumes

The Virtual I/O Server can export entire physical volumes to virtual I/O clients. Exporting physical volumes has several advantages when compared to using logical volumes as backing devices:

- Physical volumes can be exported from two or more Virtual I/O Servers, providing multipathing to virtual I/O clients.
- Physical volumes might be exported with concurrent access to more than one virtual I/O client with an appropriate concurrency management solution, such as Concurrent LVM (CLVM) and PowerHA SystemMirror.
- The code path for exporting physical volumes is shorter, which leads to better performance.
- Virtual I/O clients that are backed by external SAN devices can be moved from one Virtual I/O Server to another using LPM.
- In certain cases, existing LUNs from physical servers can be migrated into the virtual environment with the data intact. For more information, see *PowerVM Migration from Physical to Virtual Storage*, SG24-7825.

Exporting physical volumes makes administration easier, since it does not require that you manage the size by the Virtual I/O Server. The Virtual I/O Server does not allow partitioning of a single internal physical disk among multiple clients.

Rootvg disk: Moving a rootvg disk from one physical managed system to another is only supported by using LPM.

In the Virtual I/O Server, there is no need to subdivide SAN-attached disks, because storage allocation can be managed at the storage server. In the SAN environment, provision and allocate LUNs to Virtual I/O Server. Then, export them to the virtual I/O clients as physical volumes.

Cloning services: Usage of cloning services of the rootvg disk is only supported for offline backup, restore, and recovery.

reserve_policy

Whenever you need to map the same external volume to a virtual I/O client from a Virtual I/O Servers configuration, change the reserve_policy, as shown in Example 5-9. The attribute name `reserve_policy` might be called different by your storage vendor.

Example 5-9 Change the reserve_policy disk attribute

```
$ lsdev -dev hdisk12 -attr reserve_policy
value

single_path
$ chdev -dev hdisk12 -attr reserve_policy=no_reserve
hdisk12 changed
$ lsdev -dev hdisk12 -attr reserve_policy
value

no_reserve
```

Small Computer System Interface queue_depth

Increasing the value of the *queue_depth* attribute improves the throughput of the disk in some configurations. However, there are several other factors that must be considered. These factors include the value of queue_depth for all of the physical storage devices on the Virtual I/O Server being used as a virtual target device by the disk instance on the client partition. It also includes the maximum transfer size for the virtual Small Computer System Interface (SCSI) client adapter instance that is the parent device for the disk instance.

For more information about tuning the SCSI queue_depth parameter, see *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590.

Exporting virtual tape and optical devices

Virtual tape devices are assigned and operated similarly to virtual optical devices. Only one virtual I/O client can have access at a time. It is a best practice to have such devices attached to a Virtual I/O Server, instead of moving the physical parent adapter to a single client partition.

When internal tapes and optical devices are physically on the same controller as the boot disks of the Virtual I/O Server, it is a best practice to map them to a virtual host adapter. Then, use dynamic logical partitioning to assign this virtual host adapter to a client partition.

5.2.4 Configuring the Virtual I/O client with virtual SCSI

This section covers best practices to configure the virtual I/O clients. The best practices include disk and adapter tuning parameters and implementing disk load balancing between dual Virtual I/O Servers to obtain the best performance.

Assign path priorities on virtual I/O clients

In a dual Virtual I/O Server configuration, with virtual I/O clients that run AIX, the system administrator must manually configure the load balancing of the SAN traffic.

If you have a single LUN attached to two Virtual I/O Servers, use the **`chpath`** command to raise the priority of one virtual SCSI path over the other. This command is required because the MPIO algorithm only supports the failover method on the virtual I/O clients.

Prioritizing the communication paths is important to optimize and maximize the utilization of all available Virtual I/O Servers and SAN fabric bandwidth. This prioritization is also important to distribute the SAN traffic across all available Virtual I/O Server paths.

Example 5-10 on page 80 illustrates how to use the **`lspath`** command to verify the priority attribute for `hdisk0` on path `vscsi0`. Repeat the same command for `vscsi1`. If `vscsi0` and `vscsi1` have the same priority, the virtual I/O client always uses the first path that is listed in the output of **`lspath`**.

Example 5-10 Listing the hdisk0 and vscsi0 path attributes

```
# lspath -l hdisk0
Enabled hdisk0 vscsi0
Enabled hdisk0 vscsi1
# lspath -AHE -l hdisk0 -p vscsi0
attribute value description user_settable
priority  1     Priority    True
```

Example 5-11 shows how to set the path priority. This example sets `vscsi0` to the lowest priority path. When the setting of the path priority is completed, all new I/O requests use `vscsi1`, which in this case is represented by Virtual I/O Server 2. The **`chpath`** command does not require a reboot and the changes take effect immediately.

Example 5-11 Changing the vscsi0 priority for hdisk0

```
# chpath -l hdisk0 -p vscsi0 -a priority=2
```

IBM i

For IBM i, best practice is to use dual or multiple Virtual I/O Servers. Multipathing capabilities are available because of IBM i V6.1.1, which provides redundancy across multiple Virtual I/O Servers. With IBM i V7.1TR 2 or later, the IBM i multipathing algorithm is enhanced from round-robin to dynamic load balancing. With this enhancement, IBM i is able to use paths to optimize resource utilization and performance.

Virtual Small Computer System Interface tuning for AIX

When you use virtual SCSI adapters, there are some tuning parameters to maximize the performance of the virtual adapters in a virtualized environment.

The Virtual I/O Server and virtual I/O clients contain general tuning parameters that might need to be adjusted to obtain better performance. Before you make changes to any attribute, consult your storage vendor documentation.

> **Recommended values:** Remember that default values of the tuning parameters are the recommended values in most cases.

Table 5-2 on page 81 shows the recommended settings for virtual I/O clients when you use virtual SCSI disks.

Table 5-2 Recommended settings for virtual I/O client virtual SCSI disks

Parameter	Considerations and recommendations
`algorithm`	Recommended value: **fail_over** Sets how the I/Os are balanced over multiple paths to the SAN storage. A virtual SCSI disk only supports **fail_over** mode.
`hcheck_cmd`	Recommended value: **test_unit_rdy** or **inquiry**. Used to determine if a device is ready to transfer data. Change to **inquiry** only if you have reservation locks on your disks. In other cases, use the default value **test_unit_rdy**.
`hcheck_interval`	Recommended value: `60` Interval in seconds between health check polls to the disk. By default, the `hcheck_interval` is disabled. The minimum recommended value is `60` and must be configured on both the Virtual I/O Server and the virtual I/O client. The `hcheck_interval` value should not be lower than the `rw_timeout` that is configured on the Virtual I/O Server physical volume. If there was a problem with communicating to the storage, the disk driver on the Virtual I/O Server would not notice until the I/O requests times out (`rw_timeout` value). Therefore, you would not want to send a path health check before this time frame.
`hcheck_mode`	Recommended value: **nonactive** Determines which paths are checked when the health check capability is used. The recommendation is to only check paths with no active I/O.
`max_transfer`	Recommended value: Same value as in the Virtual I/O Server. The maximum amount of data that can be transferred to the disk in a single I/O operation. See the documentation from your storage vendor before you set this parameter.
`queue_depth`	Recommended value: Same value as in the Virtual I/O Server. The number of concurrent outstanding I/O requests that can be queued on the disk. See the documentation from your storage vendor before you set this parameter.
`reserve_policy`	Recommended value: **no_reserve** Provides support for applications that are able to use the SCSI-2 reserve functions.

Table 5-3 shows the recommended settings for virtual SCSI adapters in a virtual I/O client.

Table 5-3 Recommended settings for virtual I/O client virtual SCSI adapters

Parameter	Considerations and recommendations
`vscsi_err_recov`	Recommended value: **delayed_fail** or **fast_fail** Fast I/O failure might be desirable in situations where multipathing software is being used. The suggested value for this attribute is **fast_fail** when you are using a dual or multiple Virtual I/O Server configuration**.** Using fast fail, the virtual I/O client adapter might decrease the I/O fail times because of link loss between the storage device and switch. This value allows for a faster failover to alternate paths. In a single path configuration, especially configurations with a single path to a paging device, the default **delayed_fail** setting is suggested.
`vscsi_path_to`	Recommended value: `30` The virtual SCSI client adapter path timeout feature allows the client adapter to detect whether the Virtual I/O Server is not responding to I/O requests.

5.3 Shared Storage Pools

A Shared Storage Pool (SSP) allows up to four Virtual I/O Servers. The servers can be from the same frame or across different frames and can have a common clustered pool of Logical Unit Numbers (LUNs). The pool of LUNs can be used to provide storage to their virtual clients.

5.3.1 Requirements per Shared Storage Pool node

The minimum system requirements per SSP node are shown in Table 5-4.

Table 5-4 Shared Storage Pools minimum requirements per node

Category	Requirement per SSP node
Processor	1 virtual and 1 physical CPU of entitlement
Memory	4 GB
Repository storage	1 Fibre Channel attached disk of 10 GBs in size. All storage needs to be allocated on hardware RAIDed storage for redundancy.

Category	Requirement per SSP node
Data storage	1 Fibre Channel attached disk of 10 GBs in size. All storage needs to be allocated on hardware RAIDed storage for redundancy.

5.3.2 Shared Storage Pools specifications

The minimum and maximum options for SSPs are listed in Table 5-5.

Table 5-5 Shared Storage Pools minimums and maximums

Feature	Minimum	Maximum
Number of Virtual I/O Server Nodes in Cluster	1	4
Number of Physical Disks in Pool	1	256
Number of Virtual Disks (LUNs) Mappings in Pool	1	1024
Number of Client LPARs per Virtual I/O Server node	1	40
Capacity of Physical Disks in Pool	5 GB	4 TB
Storage Capacity of Storage Pool	10 GB	128 TB
Capacity of a Virtual Disk (LU) in Pool	1 GB	4 TB
Number of Repository Disks	1	1

5.3.3 When to use Shared Storage Pools

Shared Storage Pools (SSPs) provide increased flexibility for companies where system administrators constantly need to assign new storage LUNs to virtual clients. In this case, the SAN administrator provides a large portion of storage that the system administrator assigns to the SSP and distributes it to the virtual clients, as needed. This flexibility eliminates the need to contact the SAN administrator each time a new LUN is needed.

SSPs also offer various features, such as linked clones, snapshots and roll-back, and rapid provisioning. You can also easily perform an LPM operation within the frames that belong to the SSP cluster.

Starting with VIOS 2.2 FP26, non-disruptive software upgrades for applying service packs for the VIOS removed the requirement that client partitions using an SSP also be shutdown.

5.3.4 Creating the Shared Storage Pools

Creating the Shared Storage Pools (SSPs) is done by using the **`cluster`** command, and is covered in *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590.

This is a simple process that requires only a few commands. Consider the following best practices when you create SSPs:

- Create the cluster by using a single Virtual I/O Server, first.
- Add more Virtual I/O Servers as nodes to the cluster, one at a time.
- Assign storage to the virtual I/O clients.

Example 5-12 shows the process of creating the cluster, adding additional nodes, and assigning storage to the virtual I/O client.

Example 5-12 Creating the cluster, adding additional nodes, and assigning storage to the virtual I/O client

```
$ cluster -create -clustername ClusterA -repopvs hdisk1 -spname \ > StorageA -sppvs
hdisk2 hdisk3 hdisk4 -hostname vios1.itso.ibm.com
Cluster ClusterA has been created successfully.

$ cluster -addnode -clustername ClusterA -hostname vios2.itso.ibm.com
Partition vios2.itso.ibm.com has been added to the ClusterA cluster.

$ mkbdsp -clustername ClusterA -sp StorageA 10G \
> -bd aix01_datavg -vadapter vhost0
Lu Name:aix01_datavg
Lu Udid:84061064c7c3b25e2b4404568c2fcbf0

Assigning file "aix01_datavg" as a backing device.
VTD:vtscsi0
```

5.3.5 SAN storage considerations

First, plan which disks are used for the cluster repository, and which are used for the virtual I/O clients. Remember that you need a minimum of one for data and one for the repository. Then, verify that the SAN disks are present across each Virtual I/O Server that participates in the cluster.

Before a disk can be used for either a repository or storage pool disk, it is a good idea to run the **prepdev** command to determine whether the disk is already assigned or in use. Example 5-13 shows the **prepdev** command that is being run against a disk that already belongs to an SSP cluster, and the output of the recommendations before proceeding.

Example 5-13 The **prepdev** *command to prepare the disk*

```
$ prepdev -dev hdisk10
WARNING!

The VIOS has detected that this physical volume is
currently in use. Data will be lost and cannot be
undone when destructive actions are taken.
These actions should only be done after confirming
that the current physical volume usage and data are
no longer needed.

The VIOS detected that this device is a cluster disk.
Destructive action: Remove the physical volume from
the cluster by running the following command:
cleandisk -s hdisk#
```

Verifying a disk belongs to a Shared Storage Pool

Disks which make up the Shared Storage Pool (SSP) repository are placed into a volume group named `caavg_private`. However, disks which are assigned to the storage pool for virtual I/O client use, when viewed by using the **lspv** command, look to be free.

Viewing the cluster configuration with the **lscluster -c** command shows the disks that are assigned to the storage pool. Example 5-14 on page 86 shows the disks which are in the storage pool.

Example 5-14 Highlighted disks which are in the storage pool

```
$ lspv
NAME            PVID                                VG              STATUS
hdisk0          00f61ab26fed4d32                    rootvg          active
hdisk1          00f61aa6c31760cc                    caavg_private   active
hdisk2          00f61aa6c31771da                    None
hdisk3          00f61aa6c3178248                    None
hdisk4          00f61aa6c31794bf                    None
$ lscluster -c | grep disk
Number of disks in cluster = 3
         for disk hdisk4 UUID = 4aeefba9-22bc-cc8a-c821-fe11d01a5db1 cluster_major =
0 cluster_minor = 3
         for disk hdisk2 UUID = c1d55698-b7c4-b7b6-6489-c6f5c203fdad cluster_major =
0 cluster_minor = 2
         for disk hdisk3 UUID = 3b9e4678-45a2-a615-b8eb-853fd0edd715 cluster_major =
0 cluster_minor = 1
```

The `lspv` commands: Using the **`lspv -free`** or **`lspv -avail`** commands does not display the storage pool disks.

Thin or thick provisioning

Thin provisioning is a method of allocating more disk space than is physically available. Unlike thick provisioning, which allocates all the blocks up front, thin provisioning blocks are only allocated as they are written.

Best practice is to allocate storage to virtual I/O clients by using thick provisioning. However, this allocation is dependent on having sufficient capacity in the storage pool. If thin provisioning is being used, ensure that you have a reliable method of monitoring the storage utilization. Best practice is to configure threshold warnings by using the **`alert`** command, as shown in 5.3.6, “Monitoring storage pool capacity” on page 87.

Other storage considerations

Be aware of the following other storage considerations:

- Uninterrupted access to the repository disk is required for operation.
- Physical disks in the SAN storage subsystem that are assigned to the SSD, cannot be resized.
- Virtual SCSI devices that are provisioned from the SSP might drive higher processor utilization than the classic virtual SCSI devices.

- Using SCSI reservations (SCSI Reserve/Release and SCSI-3 Reserve) for fencing physical disks in the SSP, is not supported.
- High availability SAN solutions can be used to mitigate outages.

5.3.6 Monitoring storage pool capacity

Over committing the storage pool is a risk. You can monitor the free percentage threshold by using the **alert** command. A threshold of 35% is configured by default, and when triggered, will log in an entry to the error log. The default threshold is sufficient, but evaluate and configure as appropriate for your environment.

Example 5-15 shows how to list and change the threshold and the error log entry when the threshold is triggered.

Example 5-15 Using the **alert** *command to configure threshold warnings*

```
$ alert -list -clustername ClusterA -spname StorageA
PoolName:         StorageA
PoolID:           FFFFFFFFAC10161E000000004FCFA454
ThresholdPercent: 35
$ alert -set -clustername ClusterA -spname StorageA -type threshold
-value 25
$ alert -list -clustername ClusterA -spname StorageA
PoolName:         StorageA
PoolID:           FFFFFFFFAC10161E000000004FCFA454
ThresholdPercent: 25

$ errlog
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
0FD4CF1A   0606145212 I O VIOD_POOL      Informational Message
```

Available space: The `ThresholdPercent` value is the percentage of free space that is available in the storage pool.

5.3.7 Network considerations

The most important thing to remember about the network is that the hostname/IP address that is used to create the storage pool resolves to the fully qualified domain name. As an example, if you run the **hostname** command, it reports the system **hostname** as in `mydivision.mycompany.com`. If the **hostname** resolves only to the short name, then the **CLMD** command will fail.

Other considerations

Remember the following additional networking considerations:

- Uninterrupted network connectivity is required for operation. That is, the network interface that is used for SSP configuration must be on a highly reliable network, which is not congested.
- Changing the hostname/IP address for a system is not supported when it is configured in an SSP.
- Only compliant with Internet Protocol Version 4 (IPv4).
- IPv6 and VLAN tagging (IEEE 802.1Q) support for intermodal shared storage pools communication is supported.
- An SSP configuration configures the TCP/IP resolver routine for name resolution to resolve host names locally first, and then to use the Domain Name System (DNS). For step-by-step instructions, see the TCP/IP name resolution documentation in the AIX Information Center:

  http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp
- When you restore the Virtual I/O Server LPAR configuration from a **viosbr** backup, all network devices and configurations must be restored before SSP configurations are restored.
- The forward and reverse lookup resolves to the IP address/hostname that is used for the SSP configuration.
- It is suggested that the Virtual I/O Servers that are part of the SSP configuration, keep their clocks synchronized.

5.4 N-Port ID Virtualization

This section covers the benefits and best practices for choosing N-Port ID Virtualization (NPIV) as the method for virtual I/O clients to access SAN storage.

For more information about NPIV, see *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590.

5.4.1 When to use N-Port ID Virtualization

N-Port ID Virtualization (NPIV) is now the preferred method of providing virtual storage to virtual I/O clients, whenever a SAN infrastructure is available. The main advantage for selecting NPIV, compared to a virtual SCSI, is that the Virtual I/O Server is only used as a pass through to the virtual I/O client virtual

Fibre Channel adapters. Therefore, the storage is mapped directly to the virtual I/O client, with the storage allocation managed in the SAN. This strategy simplifies storage mapping at the Virtual I/O Server.

Consider the following additional benefits of NPIV:

- Provides storage resources to the virtual I/O client, just as they would be with actual physical Fibre Channel adapters.
- Multipathing takes place at the virtual I/O client.
- Allows virtual I/O clients to handle persistent reservations, which are useful in high availability cluster solutions, such as PowerHA SystemMirror.
- Makes Virtual I/O Server maintenance easier because there is no need for multipathing software.
- Is a preferred method for LPM operations.
- Using NPIV on virtual clients that run IBM i is a best practice, primarily because of performance reasons. The amount of work that the Virtual I/O Server needs to do with NPIV is less involved than with virtual SCSI.

5.4.2 Configuring the Virtual I/O Server with N-Port ID Virtualization

There are many aspects to consider when you configure the Virtual I/O Server with N-Port ID Virtualization (NPIV). Aspects to consider include LUN naming conventions, SAN infrastructure, redundancy at the Virtual I/O Server and virtual I/O client, and Fibre Channel adapter attributes.

Although the naming convention for NPIV is not as important as it is for virtual SCSI, you need to pay attention to the same naming considerations for LUNs. In this case, the naming considerations are likely to be the same as for regular non-virtualized logical partitions. An example of LUN naming in provided in, “Naming convention for external physical volumes” on page 75.

The most important thing for NPIV is to track the relationship between the virtual and physical Fibre Channel adapters. Document and maintain this relationship in the spreadsheet that you created during the planning session, described in 1.3, “Planning your Virtual I/O Server environment” on page 6.

Worldwide port names: If you plan to use LPM, remember that both the active and inactive worldwide port names (WWPNs) of a virtual Fibre Channel adapter need to be zoned from the switches. The names also need to be mapped from the storage array.

Follow these best practices for better availability, performance, and redundancy:

- Use NPIV in dual or multiple Virtual I/O Server configurations.
- Have Virtual I/O Server physical Fibre Channel adapters that split across separate system PCI busses, when possible.
- Operate at the highest possible speed on the SAN switches.
- Have the physical Fibre Channel adapter ports connected to different switches in the fabric, even in dual Virtual I/O Server configurations.
- Although it is supported to connect storage and tape libraries through NPIV by using the same Fibre Channel adapter, it is a best practice to separate them among different adapters. Separation is suggested because the disk and tape traffic have different performance characteristics and error recovery scenarios.

Configuring redundancy with N-Port ID Virtualization

As in any SAN configuration, redundancy is crucial. Therefore, avoid single points of failure. Multipathing with a dual Virtual I/O Server configuration provides the level of redundancy that is required, better performance, and flexibility for the maintenance of the Virtual I/O Servers.

Figure 5-4 on page 91 shows a dual Virtual I/O Server configuration. In this setup, virtual I/O clients avoid the loss of access to SAN storage in the event of a SAN switch, Virtual I/O Server, or physical link failure.

Each virtual I/O client has a virtual Fibre Channel adapter that is mapped to each Virtual I/O Server. Each Virtual I/O Server has a redundant link to each SAN switch.

Figure 5-4 on page 91 shows an NPIV configuration with dual Virtual I/O Servers.

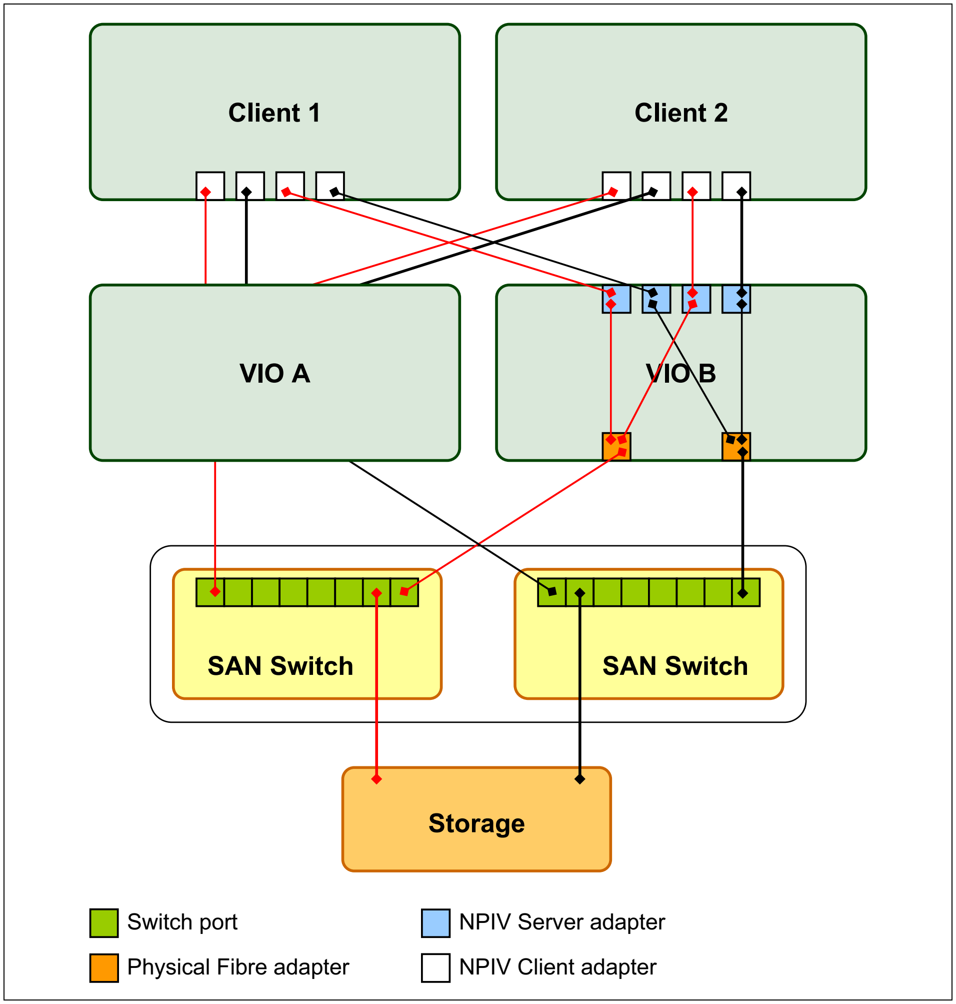


Figure 5-4 NPIV configuration with dual Virtual I/O Servers

5.4.3 Configuring the virtual I/O client with N-Port ID Virtualization

The level of configuration that is required for N-Port ID Virtualization (NPIV) in the virtual I/O client can be compared to any non-virtual logical partition. The current versions of AIX set the values of the `fc_err_recov` to `fast_fail`, by default. Confirm these values are set as shown in Example 5-16 on page 92. Change this value by using the **`chdev`** command, if needed. Dynamic tracking, **`dyntrk`**, is always enabled for virtual Fibre Channel client adapters and cannot be disabled. Dynamic tracking is enabled for the express purpose of recovering from reconfiguration events in the SAN. These events are likely to occur during LPM operations, and it is a good idea for **dyntrk** to always be set to *yes*.

Example 5-16 shows how to change values on a Fibre Channel adapter.

NPIV support: Most of the storage vendors support NPIV. Check what their requirements are to support the operating system of your virtual I/O client.

Example 5-16 Changing attributes on a Fibre Channel adapter on a virtual I/O client

```
# lsattr -El fcs0
intr_priority 3        Interrupt priority                 False
lg_term_dma   0x800000 Long term DMA                      True
max_xfer_size 0x100000 Maximum Transfer Size              True
num_cmd_elems 200      Maximum Number of COMMAND Elements True
sw_fc_class   2        FC Class for Fabric                True
# lsattr -El fscsi0
attach       none      How this adapter is CONNECTED         False
dyntrk       yes       Dynamic Tracking of FC Devices        True
fc_err_recov fast_fail FC Fabric Event Error RECOVERY Policy True
scsi_id                Adapter SCSI ID                       False
sw_fc_class  3         FC Class for Fabric                   True
```

Adopt the following best practices for virtual I/O clients that use NPIV:

- Do not change the NPIV `num_cmd_elem` and `max_xfer_size` values to be higher than the Virtual I/O Server physical adapter values. The virtual I/O client might not be able to configure new devices and might fail to boot.
- Configure a reasonable number of paths, such as 4 - 8. An excessive number of paths can increase error recovery, boot time, and the time it takes to run **cfgmgr**. This configuration can be done by zoning on SAN switches and by using LUN masking at the storage arrays.
- Balance virtual I/O client workloads across multiple physical adapters in the Virtual I/O Server.
- To avoid boot issues, balance the paths to the booting devices with **bosboot**, as documented at the following website:

http://www.ibm.com/support/docview.wss?uid=isg3T1012688

Adding a virtual Fibre Channel adapter: If you need to add a virtual Fibre Channel (FC) adapter with a dynamic logical partition operation, save the current logical partition configuration afterward. For more information, see section 3.2, “Dynamic logical partition operations” on page 42.

6

Performance monitoring

Monitoring the virtual I/O environment is essential. This process includes monitoring the critical parts of the operating system that are running on the virtual I/O clients, such as memory, processors, network, and storage. This chapter highlights best practices for monitoring the virtual environment.

© Copyright IBM Corp. 2012. All rights reserved.

6.1 Measuring Virtual I/O Server performance

Measuring the performance of a Virtual I/O Server can be generally divided into categories:

- Short-term measuring is used in testing, sizing, or troubleshooting scenarios.
- Long-term measuring is used as capacity management input because it includes performance trends or changes in workload.

6.1.1 Measuring short-term performance

On the Virtual I/O Server, use the following tools for short-term performance measurements:

viostat Reports processor statistics, and I/O statistics for the entire system, adapters, tty devices, disks, and optical devices.

netstat Reports network statistics.

vmstat Reports statistics about kernel threads, virtual memory, disks, traps, and processor activity.

topas Reports selected local system statistics, such as processor, network, processes, and memory.

seastat Reports Shared Ethernet Adapter (SEA) statistics, and generates a report for each client.

svmon Reports a snapshot of virtual memory.

fcstat Reports the statistics that are gathered by the specific Fibre Channel device driver.

If you start measuring short-term performance on the Virtual I/O Server and you do not have a specific target, such as network degradation, start with the **topas** command. The **topas** command displays local system statistics, such as system resources and Virtual I/O Server SEA statistics.

The **viostat**, **netstat**, **vmstat, svmon**, **t**seastat, and **fcstat** commands, provide more detailed output information than **topas**. Document the output of these commands and the time that they were run because it is valuable information if you need to research a performance bottleneck.

In Virtual I/O Server 2.1, the nmon functionality is integrated within the **topas** command. You can start the **topas** command and switch between the two modes by typing ~ (tilde).

Example 6-1 shows the main functions that are used in the **topas_nmon** command. You can export the environment variable nmon to start **topas_nmon** the same way every time.

Example 6-1 Display of nmon interactive mode commands

```
·h = Help information     q = Quit nmon             O = reset peak counts     ·
·+ = double refresh time  - = half refresh          r = ResourcesCPU/HW/MHz/AIX·
·c = CPU by processor     C=upto 128 CPUs           p = LPAR Stats (if LPAR)   ·
·l = CPU avg longer term  k = Kernel Internal       # = PhysicalCPU if SPLPAR  ·
·m = Memory & Paging      M = Multiple Page Sizes  P = Paging Space            ·
·d = DiskI/O Graphs       D = DiskIO +Service times o = Disks %Busy Map        ·
·a = Disk Adapter         e = ESS vpath stats       V = Volume Group stats     ·
·^ = FC Adapter (fcstat)  O = VIOS SEA (entstat)    v = Verbose=OK/Warn/Danger ·
·n = Network stats        N=NFS stats (NN for v4)   j = JFS Usage stats        ·
·A = Async I/O Servers    w = see AIX wait procs   "="= Net/Disk KB<-->MB      ·
·b = black&white mode     g = User-Defined-Disk-Groups (see cmdline -g)        ·
·t = Top-Process --->     1=basic 2=CPU-Use 3=CPU(default) 4=Size 5=Disk-I/O   ·
·u = Top+cmd arguments    U = Top+WLM Classes       . = only busy disks & procs·
·W = WLM Section          S = WLM SubClasses                                   ·
·[ = Start ODR            ] = Stop ODR                                         ·
·~ = Switch to topas screen
```

Example 6-2 shows the nmon variable that is defined to start **topas_nmon** with options to monitor processor, memory, paging, disks, and service times.

Example 6-2 Exporting environment variable in ksh shell to nmon

```
$ export NMON=cmD
$ topas_nmon
```

Monitoring processor and multiple Shared Processor Pools

To monitor the utilization of multiple Shared Processor Pools (SPPs) in a Virtual I/O Server, enable the partition profile to collect performance information. To enable the collection of cross partition processor performance data, use the Hardware Management Console (HMC) to open the partition profile properties. Then choose the hardware tab and select the check box for **Allow performance information collection**.

After the activation of the performance information collection, you can use the `topas -fullscreen lpar` command in the Virtual I/O Server to determine whether the processor resources are optimally used. Two important fields in the command output, `Psize` and `app`, are described here:

`Psize` Shows the number of online physical processors in the shared pool.

`app` Shows the available physical processors in the shared pool.

Using the same command, you can see statistics for all logical processors (LCPU) in the system, such as Power Hypervisor calls, context switching, and processes waiting to run in the queue.

On an AIX virtual I/O client, you can use the **`lparstat`** command to report logical partition-related information, statistics, and Power Hypervisor information.

On an IBM i virtual I/O client, data is collected by the collection services in the `QAPMLPARH` table. The data can be displayed either as text through SQL, or graphically through IBM System Director Navigator for i.

For more information, see *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590.

Disk monitoring

A good start to monitor the disk activity in the Virtual I/O Server is to use the **`viostat`** command.

You can get more specialized output from the **`viostat`** command. Example 6-3 on page 97 shows monitoring of `vhost1` and `hdisk4`. It is not recommended to use this output for long-term measuring because it requires a significant amount of disk space to store the data.

Example 6-3 on page 97 shows the extended disk output by using the **`viostat`** command.

Example 6-3 Extended disk output of hdisk4 and vhost1 using **`viostat`**

```
$ viostat -adapter 1 1 | grep -p vhost1 | head -2 ; viostat -extdisk hdisk4 1 1
Vadapter:                  Kbps      tps     bkread    bkwrtn
vhost1                   88064.0    688.0      344.0     344.0
System configuration: lcpu=16 drives=12 paths=17 vdisks=41

hdisk4         xfer:  %tm_act      bps      tps      bread      bwrtn
                       100.0     90.2M   344.0        0.0       90.2M
               read:      rps  avgserv  minserv  maxserv   timeouts      fails
                         0.0      0.0      0.0      0.0           0          0
              write:      wps  avgserv  minserv  maxserv   timeouts      fails
                       344.0      7.1      5.3     13.4           0          0
              queue:  avgtime  mintime  maxtime  avgwqsz    avgsqsz     sqfull
                         0.0      0.0      0.0      0.0        2.0         0.0
```

6.1.2 Network and Shared Ethernet Adapter monitoring

To monitor network and Shared Ethernet Adapters (SEAs), use the **`entstat`**, **`netstat`**, and **`seastat`** commands.

`netstat` *command*

The Virtual I/O Server **`netstat`** command provides performance data and also provides network information, such as routing table and network data. To show performance-related data, use the command with an interval, as shown in Example 6-4. Note, that you must stop the command with a Ctrl+C. This setting makes it more difficult to use the **`netstat`** command in a long-term measuring script because you specify only an interval, and not a count.

Example 6-4 Output from the **`netstat`** *command*

```
$ netstat 1
    input   (en11)     output           input   (Total)    output
 packets  errs  packets  errs colls  packets  errs  packets  errs colls
43424006     0    66342     0     0 43594413     0   236749     0     0
     142     0        3     0     0      142     0        3     0     0
     131     0        1     0     0      131     0        1     0     0
     145     0        1     0     0      145     0        1     0     0
     143     0        1     0     0      143     0        1     0     0
     139     0        1     0     0      139     0        1     0     0
     137     0        1     0     0      137     0        1     0     0
```

entstat *command*

The **entstat** command displays the statistics that are gathered by the specified Ethernet device driver. Use the **-all** flag to display all the statistics, including the device-specific statistics.

On the Virtual I/O Server, use the **entstat** command to check the status and priority of the SEA. Example 6-5 shows which SEA has the highest priority. This example uses a dual Virtual I/O Server configuration with SEAs that use the failover mode.

Example 6-5 Output of **entstat** *on SEA*

```
$ entstat -all ent9 | grep -i priority
    Priority: 1
  Priority: 1  Active: True
  Priority: 2  Active: True
```

seastat *command*

By using the **seastat** command, you can generate a report for each client to view the SEA statistics. Before you use the **seastat** command, enable `accounting` on the SEA. Example 6-6 demonstrates how to enable this option.

Example 6-6 Enabling accounting on the Shared Ethernet Adapter

```
$ lsdev -dev ent11 -attr | grep accounting
accounting    disabled Enable per-client accounting of network statistics True
$ chdev -dev ent11 -attr accounting=enabled
ent11 changed
```

With accounting enabled, the SEA tracks the MAC address of all the packets it receives from the virtual I/O clients, increment packets, and byte counts for each virtual I/O client, independently. Example 6-7 shows an output of the packets received and transmitted in a SEA, filtered by IP address.

Example 6-7 **seastat** *statistics that are filtered by IP address*

```
$ seastat -d ent9 -s ip=172.16.22.34
=======================================================================
Advanced Statistics for SEA
Device Name: ent9
=======================================================================
MAC: 22:5C:2B:95:67:04
----------------------
VLAN: None
VLAN Priority: None
```

```
IP: 172.16.22.34
Transmit Statistics:                   Receive Statistics:
--------------------                   -------------------
Packets: 29125                         Packets: 2946870
Bytes: 3745941                         Bytes: 184772445
=======================================================================
```

Fibre Channel monitoring

The **fcstat** command provides information about the adapter queue and resource use. This command is useful to determine whether modifying `max_xfer_size` or `num_cmd_elems` can increase performance. Example 6-8 shows some typical values of a Fibre Channel adapter.

Example 6-8 Example use of the **fcstat** *command*

```
$ fcstat fcs0 | grep -E "Count|Information"
LIP Count: 0
NOS Count: 0
Link Failure Count: 3
Loss of Sync Count: 261
Primitive Seq Protocol Error Count: 0
Invalid Tx Word Count: 190
Invalid CRC Count: 0
IP over FC Adapter Driver Information
  No DMA Resource Count: 0
  No Adapter Elements Count: 0
FC SCSI Adapter Driver Information
  No DMA Resource Count: 0
  No Adapter Elements Count: 0
  No Command Resource Count: 0
```

Example 6-8 shows an example of an adapter that has sufficient values for **max_xfer_size** and **num_cmd_elems**. Non zero values indicate that I/Os are being queued at the adapter because of a lack of resources. “Fibre Channel device driver attributes” on page 68 has some considerations and tuning recommendations about the **max_xfer_size** and **num_cmd_elems** parameters**.**

Memory monitoring

A best practice to monitor the memory consumption in a Virtual I/O Server is to use the **vmstat** and **svmon** commands.

The `vmstat` *command*

The **`vmstat`** command displays statistics about kernel threads, virtual memory, disks, traps, and processor activity. Regarding memory, at a high level, track the *active virtual memory* (AVM) column under memory, and page in (`PI`), and page out (`PO`) under the page column.

`AVM` is a metric that reflects virtual memory usage and is represented in 4K pages. You can convert from 4K pages to megabytes dividing the `AVM` value by 256. If `AVM` is higher than the amount of real memory, this metric can cause your Virtual I/O Server to page, which can be confirmed under the `PI` and `PO` columns. This usage is considered suboptimal for performance. We suggest a 95% threshold for `AVM` of the total amount of physical memory in the Virtual I/O Server.

The `svmon` *command*

The **`svmon`** command provides more in-depth memory information than the **`vmstat`** command. The displayed information does not constitute a true snapshot of memory because the `padmin` user does not have sufficient privileges. However, if role-based access control (RBAC) is activated, and the `vios.system.stat.memory` role is assigned to the `padmin` user, the user can see the same view that the root user sees.

Enable connection monitoring

In the partition profile for the Virtual I/O Servers and each virtual I/O client on the HMC, it is suggested to check the option for `Enable Connection Monitoring`. When connection monitoring is enabled, the service IBM Focal Point™ (FP) periodically tests the communication channel between the logical partitions and the HMC. If the channel fails, the service FP generates a serviceable event in the service FP log on the HMC. This step ensures that the communication channel can carry service requests from the logical partition to the HMC when needed.

If this option is not selected, the service FP still collects service request information when there are issues on the managed frame. This option only controls whether the service FP automatically tests the connection and generates a serviceable event if the channel fails.

For more information, see the *Hardware Management Console V7 Handbook*, SG24-7491.

6.1.3 Measuring long-term performance

In previous Virtual I/O Server versions, the Workload Manager-based commands were commonly used to measure long-term performance, but **`topasrec`** and **`topas_nmon`** commands are now a best practice for recordings.

The **topasrec** command generates binary reports of local recordings, Central Electronic Complex (CEC) recordings, and cluster recordings. In the Virtual I/O Server, persistent local recordings are stored in the /home/ios/perf/topas directory by default.

You can verify if the **topasrec** command is running by using the **ps** command. The output data is collected in a binary file, in the format `hostname.`*`yymmdd`*, for example, `localhost_120524.topas`. On an AIX server, this file can be converted to a nmon analyzer report by using the **topasout** command.

Another way to generate records is by using the nmon spreadsheet output format. You can start collecting performance data in the nmon format from the **cfgassist** menu by selecting **Performance** → **Topas** → **Start New Recording** → **Start Persistent local recording** → **nmon**. The reports can be customized according to your needs. By default, the files are stored in the directory `/home/ios/perf/topas` with file extension `nmon`. For an example of a recording in `nmon` format, see Example 6-9.

You must decide which approach is better in your environment. Best practice is to use only one format, binary or nmon, and to not increase the processor utilization. You can use the **cfgassist** menu by selecting **Performance** → **Topas** → **Stop Persistent Recording** to stop an unwanted recording, and to remove entries in the `/etc/inittab` file.

Example 6-9 shows an example of a recording in the `nmon` format.

Example 6-9 Example of a recording in the `nmon` *format*

```
.
. (Lines omitted for clarity)
.
ZZZZ,T0002,00:12:39,23-May-2012,,
CPU_ALL,T0002,0.12,0.43,0.00,99.45,4.00,
CPU03,T0002,0.00,7.56,0.00,92.44,
CPU02,T0002,0.00,7.14,0.00,92.86,
CPU01,T0002,0.16,10.17,0.00,89.67,
CPU00,T0002,19.81,63.11,0.01,17.08,
DISKBUSY,T0002,0.00,0.00,0.00,0.00,
DISKREAD,T0002,0.00,0.00,0.00,0.00,
DISKWRITE,T0002,0.00,0.00,4.93,0.00,
DISKXFER,T0002,0.00,0.00,0.59,0.00,
DISKSERV,T0002,0.00,0.00,0.00,0.00,
DISKWAIT,T0002,0.00,0.00,0.00,0.00,
MEMREAL,T0002,524288.00,226570.59,26.08,46.18,10.60,10.60,
MEMVIRT,T0002,393216.00,99.48,0.52,0.00,0.06,0.00,0.00,0.06,209.38,0.00,
PROC,T0002,0.00,1.00,454.05,375.93,37.79,3.49,1.14,1.14,0.40,0.00,99.00,
LAN,T0002,0.00,0.00,0.00,0.00,0.00,,0.00,0.00,0.00,0.00,0.00,0.00,,,0.00,0.00,0.00,0.
00,0.00,,0.00,0.00,0.00,0.00,0.00,0.00,,
,0.00,0.00,0.00,0.00,0.00,,0.00,0.00,0.00,0.00,0.00,0.00,,,0.00,0.00,0.00,0.00,0.00,,
0.00,0.00,0.00,0.00,0.00,0.00,,,0.00,0.0
0,0.00,0.00,0.00,,0.00,0.00,0.00,0.00,0.00,0.00,,,
IP,T0002,0.54,0.54,0.03,0.03,
TCPUDP,T0002,0.50,0.50,0.09,0.17,
FILE,T0002,0.00,23.94,0.00,112817.25,1125.65,
JFSFILE,T0002,36.61,0.08,81.23,39.24,0.27,80.93,
JFSINODE,T0002,11.46,0.00,26.58,6.43,0.01,21.06,
LPAR,T0002,0.01,1.00,4.00,16.00,1.00,128.00,0.00,0.07,0.07,1.00,
ZZZZ,T0003,00:17:39,23-May-2012,,
.
. (Lines omitted for clarity)
.
```

You can also use commands such as **viostat**, **netstat**, **vmstat**, **svmon**, fcstat, and **seastat** as long-term performance tools on the Virtual I/O Server. They can be used in a script, or started by using the **crontab** command to form customized reports.

Agents for long-term performance monitoring

Long-term performance of the Virtual I/O Server can be monitored and recorded by using other monitoring agents. In Example 6-10, the **lssvc** command lists which agents are supported. The **cfgsvc**, **startsvc**, **cstopsvc**, and **postprocesssvc** commands are used to manage all the available agents.

Example 6-10 Listing of the available agents

```
$ lssvc
ITM_premium
ITM_cec
TSM_base
ITUAM_base
TPC_data
TPC_fabric
DIRECTOR_agent
perfmgr
ipsec_tunnel
ILMT
```

For more information about the monitoring tools, see this website:

http://www.ibm.com/developerworks/wikis/display/WikiPtype/VIOS_Monitoring

HMC Utilization Data

HMC utilization data can record information about the memory and processor utilization on a managed frame at a particular time. You need to enable the HMC utilization data individually on each managed frame.

The data is collected into records called events. Events are created at the following times:

- At periodic intervals (hourly, daily, and monthly).
- When you make frame, logical partition, and configuration changes that affect resource utilization.
- When you start, shut down, and change the local time on the HMC.

HMC utilization data collection is deactivated by default, but can be enabled by selecting a managed frame on the HMC, and selecting **Operations** → **Utilization Data** → **Change Sampling Rate.** We suggest a sampling rate of 5 minutes.

You can use this data to analyze trends, and make resource adjustments. To display the utilization data with the HMC graphical user interface (GUI), select a managed frame, and then select **Operations** → **Utilization Data** → **View**.

On the HMC command line, you can use the **`lslparutil`** command to list the utilization data that is collected from a managed frame. You can create a script to collect the utilization data from the HMC, and export it to a comma-separated file. Example 6-11 shows an example of how to create a simple filter to list capped cycles, uncapped cycles, entitled cycles, and time cycles in a comma-separated format.

Example 6-11 The **`lslparutil`** *command*

```
hscroot@hmc9:~> lslparutil -m POWER7_1-SN061AA6P -r lpar --filter \
> "lpar_names=vios1a" -n 2 -F time,lpar_id,capped_cycles,\
> uncapped_cycles,entitled_cycles,time_cycles
06/08/2012
14:21:13,1,2426637226308,354316260938,58339456586721,58461063017452
06/08/2012
14:20:13,1,2425557392352,354159306792,58308640647690,58430247078431
```

Example 6-12 on page 104 shows how to calculate the processor utilization for the shared processor partition, 1 minute apart. The sampling rate is a 1-minute interval that is based on the data that is collected from Example 6-11.

Example 6-12 How to calculate processor utilization

```
Processor utilization % =
            ((capped_cycles + uncapped_cycles) / entitled_cycles) * 100
        Processor utilization % = (((2426637226308 - 2425557392352) +
(354316260938 - 354159306792)) / (58339456586721 - 58308640647690)) *
100

Processor utilization % = 4.01%

        Processor units utilized = (capped_cycles + uncapped_cycles) /
time_cycles
        Processor units utilized = ((2426637226308 - 2425557392352) +
(354316260938 - 354159306792)) / (58461063017452 - 58430247078431)

Processor units utilized = 0.04
```

Virtual I/O Server Advisor

The Virtual I/O Server Advisor polls and collects key performance metrics. It analyzes the results, provides a health check report, and then proposes changes to the environment or suggests areas to investigate further.

Best practice is to run the **`vios_advisor`** command for a 30 minute period after the Virtual I/O Server is configured and has an active workload that runs through it. This practice ensures that the Virtual I/O Server has sufficient resources and that those resources are correctly configured to handle requests during all peak periods.

Note: You can record a minumum of 5 minutes and up to 24 hours, but note that running the Advisor consumes system resources that may affect VIOS performance.

The Virtual I/O Server Advisor is a utility that is available at no charge. The tool can be downloaded from this website:

http://www.ibm.com/developerworks/wikis/display/WikiPtype/VIOS+Advisor

Important: At the time of writing, the VIOS Advisor was a separate downloadable add on. Since then, an IBM supported version is available in the perf.analysis fileset starting 2.2.2.0 VIOS level

7

Security and advanced IBM PowerVM features

This chapter provides best practices for security on your virtual environment. It also covers best practices on the advanced features that are in the IBM PowerVM Enterprise Edition, which includes Active Memory Sharing (AMS) and Live Partition Mobility (LPM).

© Copyright IBM Corp. 2012. All rights reserved.

7.1 Virtual I/O Server security

Security is vital for all businesses today, especially because most companies are moving from physical to virtual environments. This move eliminates physical barriers and adopts new computing models, such as cloud computing, where resources are shared among multiple tenants.

Section 7.1 describes best practices for security on your virtual environment.

7.1.1 IBM PowerVM Hypervisor security

The IBM PowerVM Hypervisor is secure by design, and does not have a single security vulnerability to date. It provides isolation among the dedicated and virtual partitions.

Your data in a virtual client is safe. You can share hardware resources, such as processor and memory usage, among these virtual partitions. And, you can be confident that the Power Hypervisor handles the separation of resources in a secured way.

To date, there are no reported security vulnerabilities on the PowerVM Hypervisor. However, as a best practice, ensure that your system is up-to-date on the latest firmware to incorporate any new functions, or to fix any issues.

7.1.2 Virtual I/O Server network services

After installation, the Virtual I/O Server has some services which are open and running, by default. The services in the listening open state are listed in Table 7-1.

Table 7-1 Default open ports on the Virtual I/O Server

Port number	Service	Purpose
21	FTP	Unencrypted file transfer
22	SSH	Secure Shell and file transfer
23	Telnet	Unencrypted remote login
111	rpcbind	NFS connection
657	RMC	RMC connection (used for dynamic LPAR operations)

In most cases, the Secure Shell (SSH) service for remote login and the Secure Copy Protocol (SCP) for copying files is sufficient for login and file transfer. Telnet and File Transfer Protocol (FTP) are not using encrypted communication and can be disabled.

Port 657 for Resource Monitoring and Control (RMC) must be left open if you are considering the use of dynamic logical partition operations. This port is used for the communication between the logical partition and the Hardware Management Console (HMC).

The stopping of these services can be done by using the **stopnetsvc** command.

7.1.3 Viosecure command

The **viosecure** command allows for the configuration of security hardening and firewall rules on the Virtual I/O Server. Firewall rules, in most cases, are environment-specific, and are out of the scope of this document. For more information about how to use **viosecure** to configure the firewall, see *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590

Security hardening rules

The **viosecure** command can also be used to configure security hardening rules. Users can enforce either the preconfigured security levels or choose to customize them, based on their requirements.

Best practice when you use **viosecure** is to export all rules to an XML file format and manually modify this file based on the security rule requirements. The command, as shown in Example 7-1, exports all rules which apply to the level high, to a file named viosecure.high.xml.

Example 7-1 Exporting **viosecure** *high-level rules to XML*

```
$ viosecure -level high -outfile viosecure.high.xml
```

iviosecure command rules: The full list of **viosecure** rules is large. To help editing, we suggest copying the file to your desktop, and using a text editor with XML support.

The **viosecure** command uses the same rule set definitions as **aixpert** on IBM AIX. For a full list of rules and applicable values, see this website:

http://www.ibm.com/developerworks/wikis/display/WikiPtype/aixpert

When you finish customizing the XML file, copy it back to the Virtual I/O Server and apply it using the command, as shown in Example 7-2 on page 108.

Example 7-2 Applying customized **`viosecure`** *rules*

```
$ viosecure -file viosecure.high.xml
```

`viosecure` command: We suggest running the **`viosecure`** command from a console session because some rules might result in the loss of access through the network connections.

padmin password: The `maxage` and `maxexpired` stanzas, when applied, might result in the padmin account being disabled because of the aging of the password. We suggest changing the padmin password before you run the **`viosecure`** command to prevent the account from becoming disabled.

7.2 IBM PowerSC

IBM PowerSC is a separate licensed product that can assist you in the security and compliance of your virtual environment. PowerSC enables security compliance automation and includes reporting for compliance measurement and audit.

Your business might need to adhere to specific regulatory requirements or industry data security standards, such as the Payment Card Industry (PCI) standard. If so, it is much easier for you to meet these requirements and stay compliant by using PowerSC in your virtual environment.

Table 7-2 lists the current PowerSC features and their benefits. Bookmark the following link to stay current with the new PowerSC features:

http://ibm.com/systems/power/software/security/features.html

Table 7-2 IBM PowerSC security standards

Feature	Benefits
Security and compliance automation	Reduces administration costs for complying with industry security standards.
Compliance reports	Reduces time and cost to provide security and compliance reports to auditors.
Preconfigured profiles for PCI, DOD STIG, and COBIT	Saves time, cost, and risk that is associated with deploying industry security standards.

Feature	Benefits
Trusted Boot	Reduces risk of compromised security by guaranteeing that an AIX operating system image is not inadvertently or maliciously altered.
Trusted Monitoring	Ensures high levels of trust by displaying the status of all AIX systems that participate in a trusted system configuration.
Trusted Logging	Prevents tampering or covering security issues by storing AIX virtual machine system logs securely on a central PowerVM Virtual I/O Server. Reduces backup and archive time by storing audit logs in a central location.
Trusted Network Connect And Patch Management	Ensures that site patch levels policies are adhered to in virtual workloads. Provides notification of noncompliance when back-level systems are activated.
Trusted Firewall	Improves performance and reduces network resource consumption by providing firewall services locally with the virtualization layer.

7.3 Live Partition Mobility

Live Partition Mobility (LPM) is a feature of the PowerVM Enterprise Edition that can move live AIX, Linux, and IBM i logical partitions from one frame to another, without any downtime. It can also be used to move inactive logical partitions. The mobility process transfers the virtual I/O client partition environment, including the processor state, memory, attached virtual devices, and connected users.

This topic describes the best practices when you implement and manage an LPM environment.

Implementing LPM: See *IBM PowerVM Live Partition Mobility*, SG24-7460 for information about how to implement LPM.

7.3.1 General considerations

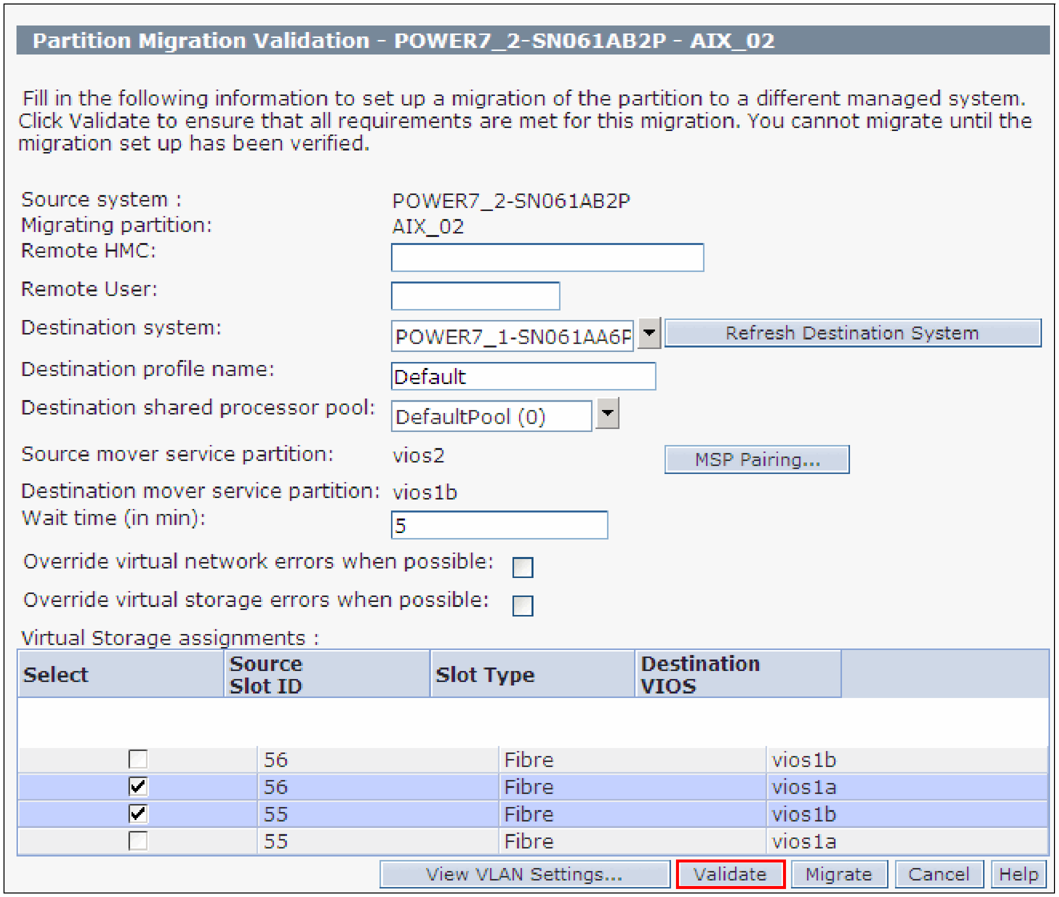
The following list provides best practice scenarios which can benefit from the use of LPM:

- Balancing the frame workload across multiple frames. For example, if you need to increase the processor or memory allocation for a partition and the frame does not have the required capacity, you can move partitions to another frame to free up resources.
- LPM offers flexibility in repair actions and updates. For example, you can migrate all the running partitions to another frame, while you perform Virtual I/O Server or frame firmware updates.
- LPM can be used to migrate partitions to a new frame when a hardware refresh is done, or to consolidate Power Systems server environments.

7.3.2 Implementing Live Partition Mobility

The following list provides best practice considerations when you implement Live Partition Mobility (LPM):

- Before you perform an LPM operation, best practice is to run the partition migration validation function on the HMC. This function checks that the frame, Virtual I/O Server, LPARs, storage, and network are ready for LPM operations. Figure 7-1 on page 111 shows the LPM validation function.



Partition Migration Validation - POWER7_2-SN061AB2P - AIX_02

Fill in the following information to set up a migration of the partition to a different managed system. Click Validate to ensure that all requirements are met for this migration. You cannot migrate until the migration set up has been verified.

Source system :	POWER7_2-SN061AB2P
Migrating partition:	AIX_02
Remote HMC:	
Remote User:	
Destination system:	POWER7_1-SN061AA6P
Destination profile name:	Default
Destination shared processor pool:	DefaultPool (0)
Source mover service partition:	vios2
Destination mover service partition:	vios1b
Wait time (in min):	5

Override virtual network errors when possible: □

Override virtual storage errors when possible: □

Virtual Storage assignments :

Select	Source Slot ID	Slot Type	Destination VIOS
□	56	Fibre	vios1b
☑	56	Fibre	vios1a
☑	55	Fibre	vios1b
□	55	Fibre	vios1a

Figure 7-1 Live Partition Mobility migration validation function

- After you perform an LPM, document all changes that are made during the operation. When you move a partition to a different frame, there might be changes to the partition ID and the adapter slot numbers that were originally configured.
- When possible, perform LPM operations outside of peak hours. LPM is more efficient when the load on the network is low.

> **LPM:** It is important to remember that LPM is not a replacement for PowerHA SystemMirror or a disaster recovery solution. LPM can move a powered-off partition, but not a crashed-kernel partition. Logical partitions cannot be migrated from failed frames.

7.3.3 Storage considerations

The following storage area network (SAN) considerations must be checked when you implement LPM:

- When you configure a virtual Small Computer System Interface (SCSI), the storage must be zoned to both the source and target Virtual I/O Servers. Also, only SAN disks are supported in LPM.
- When you use N-Port ID Virtualization (NPIV), confirm that both worldwide port names (WWPNs) on the virtual Fibre Channel adapters are zoned.
- Dedicated I/O adapters must be deallocated before migration. Optical devices in the Virtual I/O Server must not be assigned to the virtual I/O clients that are moved.
- When you use the virtual SCSI adapters, verify that the reserve attributes on the physical volumes are the same for the source and destination Virtual I/O Servers.
- When you use the virtual SCSI before you move a virtual I/O client, you can specify a new name for the virtual target device (VTD). Do this step if you want to preserve the same naming convention on the target frame. After you move the virtual I/O client, the VTD assumes the new name on the target Virtual I/O Server. Example 7-3 shows the command to specify a new VTD name.

Example 7-3 Specifying a new name for the VTD

```
VTD name, before LPM, on the source frame.

$lsmap -vadapter vhost1
SVSA            Physloc    Client                           Partition
ID
--------------- -------------------------------------------- ------
vhost1          U8233.E8B.061AA6P-V2-C302                   0x00000003

VTD                   client1_hd0
Status                Available
LUN                   0x8100000000000000
Backing device        hdisk4

$ chdev -dev client1_hd0 -attr mig_name=lpm_client1_hd0

VTD name, after LPM, on the target frame.

$lsmap -vadapter vhost1
```

```
SVSA            Physloc    Client                              Partition
ID
--------------- ------------------------------------------- ------
vhost1         U8233.E8B.061AA6P-V2-C302                      0x00000003

VTD                   lpm_client1_hd0
Status                Available
LUN                   0x8100000000000000
Backing device        hdisk4
```

7.3.4 Network considerations

The following list provides network considerations when you implement LPM:

- Shared Ethernet Adapters (SEA) must be used in an LPM environment.
- Source and target frames must be on same subnet to bridge the same Ethernet network that the mobile partitions use.
- The network throughput is important. The higher the throughput, the less time it takes to perform the LPM operation.

  For example, if you are performing an LPM operation on a virtual I/O client with 8 GB of memory:

  – A 100-MB network, sustaining a 30-MB/s throughput, takes 36 minutes to complete the LPM operation.

  – A 1-GB network, sustaining a 300-MB/s throughput, takes 3.6 minutes to complete the LPM operation.

7.4 Active Memory Sharing

IBM Active Memory Sharing (AMS) is a PowerVM advanced memory virtualization technology that provides system memory virtualization capabilities to IBM Power Systems. This technology allows multiple logical partitions to share a common pool of physical memory.

A system administrator configures a logical partition with enough memory to satisfy the workload and to avoid paging activity. However, such memory demands might be needed for only a short amount of time; for example, during workload peak times. At times, the changes in these workloads can be difficult to predict. When a system is configured with AMS, it automatically adjusts the amount of memory that is needed by the virtual client, depending on the workload.

With dedicated memory, the allocation of memory to each partition is static, as shown in Figure 7-2. The amount of memory, regardless of demand, does not change.

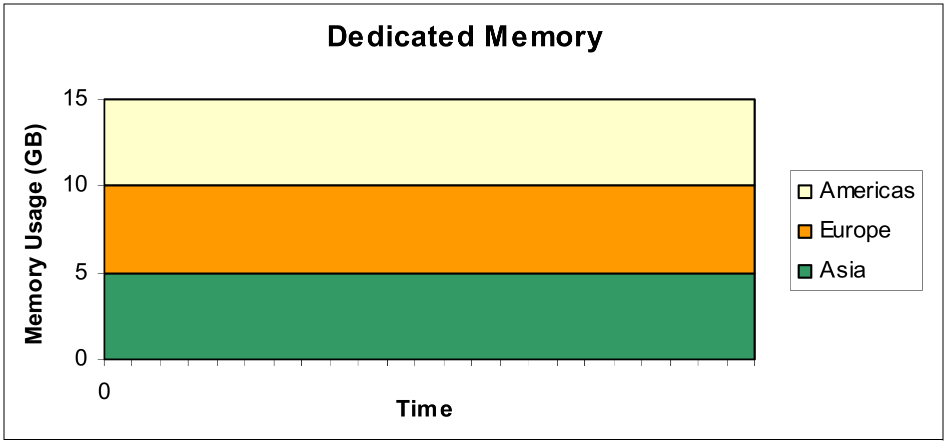


Figure 7-2 Partitions with dedicated memory

With AMS, the amount of memory that is made available to a partition, changes over the run time of the workload. The amount is based on the memory demands of the workload.

AMS is supported by AIX, IBM i, and Linux. For more information about the setup and requirements, see the IBM Redpaper™ publication, *IBM PowerVM Virtualization Active Memory Sharing*, REDP-4470, available at this website:

http://www.redbooks.ibm.com/abstracts/redp4470.html

7.4.1 When to use Active Memory Sharing

The following three examples are perfect candidates for Active Memory Sharing (AMS):

- Geographically separated workloads.
- Day and night workloads.
- Many logical partitions with sporadic use.

Geographically separated workloads

As a best practice, we suggest that you use AMS in an environment which supports workloads for different regions. A suitable environment might include geographically separated workloads where logical partition memory demands are different for day and night, or consist of many logical partitions with sporadic use.

Figure 7-3 shows three partitions with similar memory requirements. With this scenario, it is likely that the workloads will peak memory resources at different times throughout a 24 hour period. High memory requirements in America during business hours might mean that memory requirements in Asia are low.

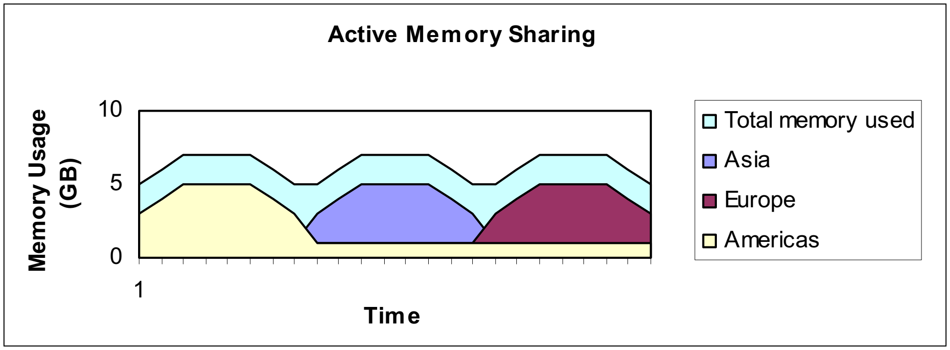


Figure 7-3 Logical partitions with shared memory that run different regions

Day and night workloads

An environment which supports day and night workloads might be for logical partitions which run applications during the day time and process batch jobs at night. Figure 7-4 shows an example where the night batch process uses less memory. The unutilizied memory can be given back to the pool for other use.

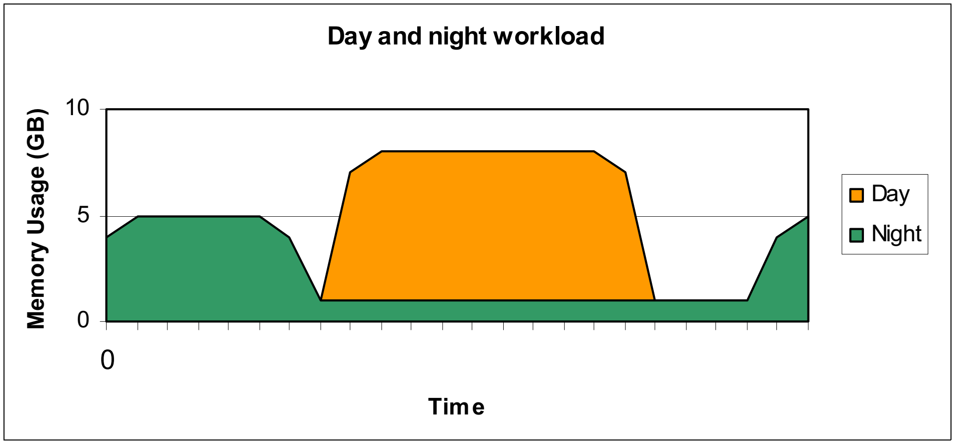


Figure 7-4 Partitions that support day and night workloads

Large number of logical partitions with sporadic use

AMS can also be used in systems that have many logical partitions, but the use of these logical partitions is intermittent, as shown in Figure 7-5.

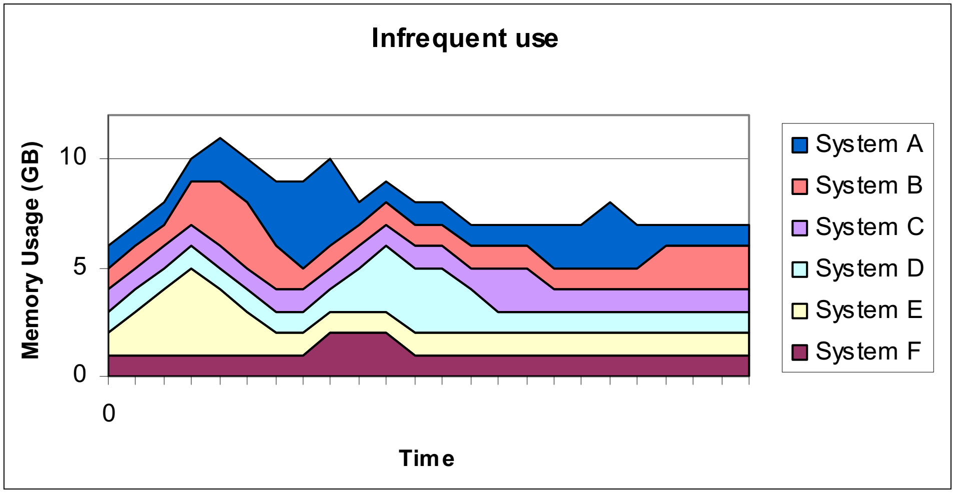


Figure 7-5 Sporadically used logical partitions

7.4.2 Implementing Active Memory Sharing

When you introduce Active Memory Sharing (AMS) into a system, it is important to plan the setup of the minimum, wanted, maximum memory, memory weighting, and the paging devices.

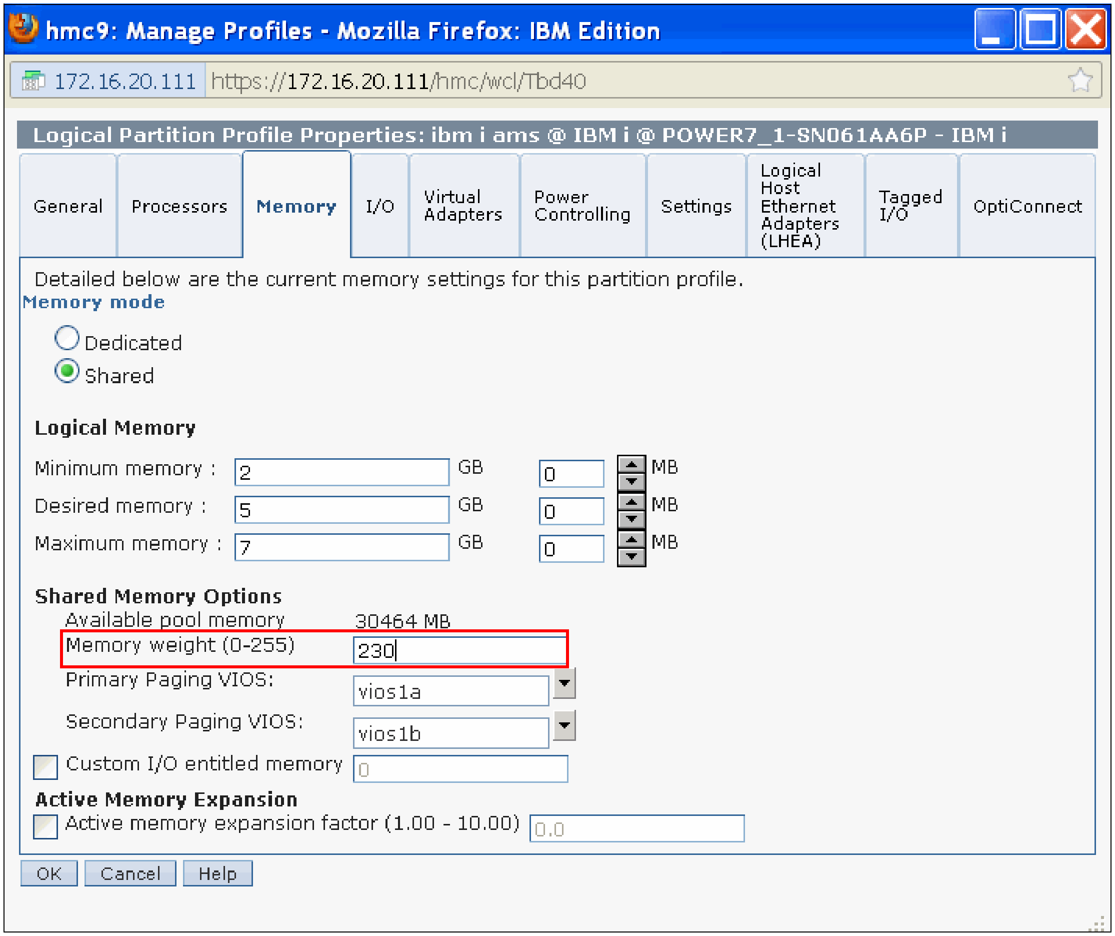
For minimum memory, in most cases, 1 - 2 GB is normally sufficient. However, the size is dependent on the server workload.

The wanted amount of memory needs to be configured to a value which is sufficient for the workload, and to avoid paging activity.

For maximum memory, add 20 - 30% to the wanted memory value.

The maximum memory value is important for two reasons. First, if the maximum memory value is set too low, and a system encounters a spike in workload (for instance, natural growth, or at the end of a financial quarter), you are able to dynamically add memory only up to the maximum value configured. Second, the maximum memory value is used by the Hypervisor to calculate and reserve the page tables. To avoid wasting memory, it is a best practice to carefully plan, and keep maximum memory values to a minimum.

Memory weight defines a factor that is used by the Power Hypervisor in determining the allocation of physical system memory. It determines which pages must be copied to the paging devices in case of physical memory over-commitment. The best practice is to set a higher weight for production systems and a lower weight for development systems. Figure 7-6 shows the weight configuration on a logical partition.



hmc9: Manage Profiles - Mozilla Firefox: IBM Edition

172.16.20.111 https://172.16.20.111/hmc/wcl/Tbd40

Logical Partition Profile Properties: ibm i ams @ IBM i @ POWER7_1-SN061AA6P - IBM i

Detailed below are the current memory settings for this partition profile.

Memory mode

○ Dedicated
● Shared

Logical Memory

Minimum memory : 2 GB 0 MB
Desired memory : 5 GB 0 MB
Maximum memory : 7 GB 0 MB

Shared Memory Options

Available pool memory 30464 MB
Memory weight (0-255) 230
Primary Paging VIOS: vios1a
Secondary Paging VIOS: vios1b
□ Custom I/O entitled memory 0

Active Memory Expansion

□ Active memory expansion factor (1.00 - 10.00) 0.0

Figure 7-6 AMS memory weight on a logical partition

Paging devices

Response from the AMS paging devices has a significant effect for clients when there is a physical memory over-commitment. You want these operations to be achieved as fast as possible. The best practice is to set the size of the paging device equal to the maximum logical memory, with an exception for IBM i. In IBM i, the paging device must be larger than the maximum memory defined in the partition profile. It requires 1 bit extra for every 16-byte page it allocates. For instance, a partition with 10 Gb of memory needs a paging device with a minimum size of 10.08 Gb defined in the partition profile.

Remember, some storage vendors do not support dynamically increasing the size of a logical unit number (LUN). If this case, the paging device needs to be removed and re-created. The following list outlines best practices to consider when you configure paging devices:

- Use physical volumes, where possible, over logical volumes.
- If you use logical volumes, use a small stripe size.
- Use thin provisioned LUNs.
- Spread the I/O load across as much of the disk subsystem as possible.
- Use a write cache, whether it is on the adapter or storage subsystem.
- Size your storage hardware according to your performance needs.
- Ensure that the PVIDs for the paging devices for physical volumes that are set up by the HMC, are cleared before use.
- If you plan to use a dual Virtual I/O Server configuration, paging devices must be provided through a SAN, and accessible from both Virtual I/O Servers.

Active Memory Expansion

If you are using AMS on AIX, best practice is to enable *Active Memory Expansion* (AME). AME allows a partition to expand its memory up to a specific factor. With AME enabled, page loaning (an AMS mechanism) tries to compress memory page content before you move it to the paging device. When compression is no longer possible, AMS moves memory pages straight to the paging device.

7.4.3 Active Memory Deduplication

Active Memory Deduplication is a PowerVM technology that minimizes the existence of identical memory pages in the main memory space. When you run workloads on traditional partitions, multiple identical data is saved across different positions in the main memory. Active Memory Deduplication can improve usage of physical memory. As a best practice, we suggest that you enable Active Memory Deduplication if you use AMS and a Virtual I/O Server that is not processor-constrained.

The Power Hypervisor reserves extra memory for page tracking and calculation of the deduplication table ratio. Therefore, it is suggested that you keep the maximum shared memory pool size parameter to a minimum. For more information about setup and configuration of Active Memory Deduplication, see *Power Systems Memory Deduplication*, REDP-4827.

Abbreviations and acronyms

ABI	application binary interface
AC	Alternating Current
ACL	access control list
AFPA	Adaptive Fast Path Architecture
AIO	Asynchronous I/O
AIX	Advanced Interactive Executive
APAR	authorized program analysis report
API	application programming interface
ARP	Address Resolution Protocol
ASMI	Advanced System Management Interface
BFF	Backup File Format
BIND	Berkeley Internet Name Domain
BIST	Built-In Self-Test
BLV	Boot Logical Volume
BOOTP	Bootstrap Protocol
BOS	Base Operating System
BSD	Berkeley Software Distribution
CA	certificate authority
CATE	Certified Advanced Technical Expert
CD	compact disc
CDE	Common Desktop Environment
CD-R	compact disc recordable
CD-ROM	compact-disc read-only memory
CEC	central electronics complex
CHRP	Common Hardware Reference Platform
CLI	command-line interface
CLVM	Concurrent LVM
CPU	central processing unit
CRC	cyclic redundancy check
CSM	Cluster Systems Management
CUoD	Capacity Upgrade on Demand
DCM	Dual Chip Module
DES	Data Encryption Standard
DGD	Dead Gateway Detection
DHCP	Dynamic Host Configuration Protocol
DLPAR	dynamic LPAR
DMA	direct memory access
DNS	Domain Name System
DRM	dynamic reconfiguration manager
DR	dynamic reconfiguration
DVD	digital versatile disc
EC	EtherChannel
ECC	error correction code
EOF	end-of-file
EPOW	emergency power-off warning
ERRM	Event Response resource manager
ESS	IBM Enterprise Storage Server®
F/C	Feature Code
FC	Fibre Channel
FC_AL	Fibre Channel Arbitrated Loop
FDX	Full Duplex

© Copyright IBM Corp. 2012. All rights reserved.

FLOP	Floating Point Operation
FRU	field-replaceable unit
FTP	File Transfer Protocol
GDPS®	IBM Geographically Dispersed Parallel Sysplex™
GID	group ID
GPFS	General Parallel File System
GUI	graphical user interface
HACMP	High Availability Cluster Multiprocessing
HBA	host bus adapter
HMC	Hardware Management Console
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
Hz	hertz
I/O	input/output
IBM	International Business Machines
ID	identifier
IDE	Integrated Device Electronics
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
IPAT	IP address takeover
IPL	initial program load
IPMP	IP Multipathing
ISV	independent software vendor
ITSO	International Technical Support Organization
IVM	Integrated Virtualization Manager
JFS	journaled file system
L1	level 1
L2	level 2
L3	level 3
LA	Link Aggregation
LACP	Link Aggregation Control Protocol
LAN	local area network
LDAP	Lightweight Directory Access Protocol
LED	light-emitting diode
LMB	Logical Memory Block
LPAR	logical partition
LPP	licensed program product
LUN	logical unit number
LV	logical volume
LVCB	Logical Volume Control Block
LVM	Logical Volume Manager
MAC	Media Access Control
Mbps	megabits per second
MBps	megabytes per second
MCM	multiple chip module
ML	Maintenance Level
MP	Multiprocessor
MPIO	Multipath I/O
MTU	maximum transmission unit
NFS	Network File System
NIB	Network Interface Backup
NIM	Network Installation Management
NIMOL	NIM on Linux
N_PORT	Node Port
NPIV	N_Port Identifier Virtualization
NVRAM	nonvolatile random access memory
ODM	Object Data Manager
OS	operating system
OSPF	Open Shortest Path First
PCI	Peripheral Component Interconnect
PCI-e	iPeripheral Component Interconnect Express

PIC	Pool Idle Count
PID	process ID
PKI	public key infrastructure
PLM	Partition Load Manager
POST	power-on self-test
POWER	Performance Optimization with Enhanced Risc (Architecture)
PPC	Physical Processor Consumption
PPFC	Physical Processor Fraction Consumed
PTF	program temporary fix
PTX	Performance Toolbox
PURR	Processor Utilization Resource Register
PV	physical volume
PVID	Port Virtual LAN Identifier
QoS	quality of service
RAID	Redundant Array of Independent Disks
RAM	random access memory
RAS	reliability, availability, and serviceability
RBAC	role-based access control
RCP	Remote Copy
RDAC	Redundant Disk Array Controller
RIO	remote input/output
RIP	Routing Information Protocol
RISC	reduced instruction-set computer
RMC	Resource Monitoring and Control
RPC	Remote Procedure Call
RPL	Remote Program Loader
RPM	Red Hat Package Manager
RSA	Rivest-Shamir-Adleman algorithm
RSCT	Reliable Scalable Cluster Technology
RSH	Remote Shell
SAN	storage area network
SCSI	Small Computer System Interface
SDD	Subsystem Device Driver
SDDPCM	Subsystem Device Driver Path Control Module
SMIT	System Management Interface Tool
SMP	symmetric multiprocessor
SMS	system management services
SMT	simultaneous multithreading
SP	Service Processor
SPOT	Shared Product Object Tree
SRC	System Resource Controller
SRN	service request number
SSA	Serial Storage Architecture
SSH	Secure Shell
SSL	Secure Sockets Layer
SUID	Set User ID
SVC	SAN Volume Controller
TCP/IP	Transmission Control Protocol/Internet Protocol
TL	Technology Level
TSA	Tivoli System Automation
UDF	Universal Disk Format
UDID	Universal Disk Identification
VIPA	virtual IP address
VG	volume group
VGDA	Volume Group Descriptor Area
VGSA	Volume Group Status Area
VLAN	virtual local area network

VP	Virtual Processor
VPD	vital product data
VPN	virtual private network
VRRP	Virtual Router Redundancy Protocol
VSD	Virtual Shared Disk
WLM	Workload Manager
WWN	worldwide name
WWPN	worldwide port name

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *Hardware Management Console V7 Handbook*, SG24-7491
- *IBM PowerVM Virtualization Active Memory Sharing*, REDP-4470
- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- *Integrated Virtualization Manager on IBM System p5*, REDP-4061
- *Power Systems Memory Deduplication*, REDP-4827
- *PowerVM Migration from Physical to Virtual Storage*, SG24-7825
- *IBM System Storage DS8000 Host Attachment and Interoperability*, SG24-8887
- *IBM Flex System p260 and p460 Planning and Implementation Guide*, SG24-7989

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources.

- The following types of documentation are located through the Internet at the following URL:

  http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp

© Copyright IBM Corp. 2012. All rights reserved.

  - User guides
  - System management guides
  - Application programmer guides
  - All commands reference volumes
  - Files reference
  - Technical reference volumes used by application programmers
- Detailed documentation about the PowerVM feature and the Virtual I/O Server:

  https://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/home.html

Online resources

These Web sites are also relevant as further information sources:

These Web sites and URLs are also relevant as further information sources:

- AIX and Linux on POWER community

  http://www-03.ibm.com/systems/p/community/
- Capacity on Demand

  http://www.ibm.com/systems/p/cod/
- IBM PowerVM

  http://www.ibm.com/systems/power/software/virtualization/index.html
- AIX 7.1 Information Center

  http://publib.boulder.ibm.com/infocenter/aix/v7r1/index.jsp
- IBM System Planning Tool

  http://www.ibm.com/servers/eserver/support/tools/systemplanningtool/
- IBM Systems Hardware Information Center

  http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp
- IBM Systems Workload Estimator

  http://www-304.ibm.com/jct01004c/systems/support/tools/estimator/index.html
- Latest *Multipath Subsystem Device Driver home page*

  http://www-1.ibm.com/support/docview.wss?uid=ssg1S4000201

- Novell SUSE LINUX Enterprise Server information

  http://www.novell.com/products/server/index.html
- SCSI T10 Technical Committee

  http://www.t10.org
- SDDPCM software download page

  http://www.ibm.com/support/docview.wss?uid=ssg1S4000201
- Service and productivity tools for Linux on POWER

  https://www14.software.ibm.com/webapp/set2/sas/f/lopdiags/home.html
- Virtual I/O Server home page

  http://www14.software.ibm.com/webapp/set2/sas/f/vios/home.html
- Virtual I/O Server fix pack download page

  http://www14.software.ibm.com/webapp/set2/sas/f/vios/download/home.html

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

© Copyright IBM Corp. 2012. All rights reserved.

IBM Redbooks

IBM PowerVM Best Practices

(0.2”spine)
0.17”<->0.473”
90<->249 pages

IBM PowerVM Best Practices

A collection of recommended practices to enhance your use of the PowerVM features

A resource to build on knowledge found in other PowerVM documents

A valuable refererence for experienced IT specialists and IT architects

This IBM Redbooks publication provides best practices for planning, installing, maintaining, and monitoring the IBM PowerVM Enterprise Edition virtualization features on IBM POWER7 processor technology-based servers.

PowerVM is a combination of hardware, PowerVM Hypervisor, and software, which includes other virtualization features, such as the Virtual I/O Server.

This publication is intended for experienced IT specialists and IT architects who want to learn about PowerVM best practices, and focuses on the following topics:

- Planning and general best practices
- Installation, migration, and configuration
- Administration and maintenance
- Storage and networking
- Performance monitoring
- Security
- PowerVM advanced features

This publication is written by a group of seven PowerVM experts from different countries around the world. These experts came together to bring their broad IT skills, depth of knowledge, and experiences from thousands of installations and configurations in different IBM client sites.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-8062-00 ISBN 0738437301